State of the Art LSA

State of the Art LSA

Editor	*Open University of the Netherlands*
Status	*Final Version*
Date	*1 December 2006*

This document was realised as part of deliverable 3.1 within the Cooper project: Iofciu, T., Zhou, X., Giesbers, B., Rusman, E., van Bruggen, J., Ceri, S. (2006). State of the Art Report in Knowledge Sharing, Recommendation and Latent Semantic Analysis. This document contains the text about LSA, plus an additional chapter on LSA methodology which had not been included in the Cooper deliverable.

Collaborative Open Environment for Project-Centred Learning

European Commission Sixth Framework Project (IST-027073)

Document Control

Title: **Final**

Author/Editor: **Open University of the Netherlands**

E-mail: **Bas.giesbers@ou.nl; Ellen.rusman@ou.nl; Jan.vanbruggen@ou.nl**

Amendment History

Version	Date	Author/Editor	Description/Comments
1	March 2006	Bas Giesbers, Ellen Rusman, Jan van Bruggen	Outline
2	March 2006	Bas Giesbers, Ellen Rusman, Jan van Bruggen	First draft
3	April 2006	Bas Giesbers, Ellen Rusman, Jan van Bruggen	Second draft
4	May 2006	Bas Giesbers, Ellen Rusman, Jan van Bruggen	Third draft, after a review of the second draft by Stefan Trausan-Matu (University Politehnica of Bucharest, Romania).
5	November 2006	Bas Giesbers, Ellen Rusman	Text changes

Legal Notices

1 INTRODUCTION 5

2 WHAT IS LATENT SEMANTIC ANALYSIS? 5

2.1 Definition 5

2.2 LSA in a nutshell 8

3 APPLICATION AREAS OF LSA 12

3.1 Document retrieval and latent semantic indexing (LSI) 12

3.1.1 Synonymy and polysemy 13

3.2 Representation of semantics and discourse processing 13

3.3 Educational applications 15

3.3.1 Intelligent tutoring systems- assessment and feedback of free text responses 16

3.3.2 Intelligent tutoring systems- selection and sequencing of instruction 16

3.3.3 Community formation and community support 19

3.3.4 Question answering 20

3.3.5 Accreditation of Prior Learning and positioning 20

3.4 Human Resource Management 21

4 IMPLEMENTATION ISSUES: CORPUS CONSTRUCTION 22

4.1.1 Corpus size 22

4.1.2 Document selection 23

4.1.3 Document size 23

4.1.4 Stemming 24

4.1.5 Stopping 24

5 METHODOLOGICAL CONSIDERATIONS 24

5.1 Determining the number of singular values 24

5.2 Weighting: influencing vector length 26

5.3 Measure of similarity 26

5.4 Queries 26

5.5 Updating and folding 27

6 EVALUATION 27

6.1 Strengths of LSA 27

6.1.1 Strengths through mathematical representation 27

6.1.2 Strengths through representation and reduction of complexity of concepts 28

6.2 Weaknesses of LSA 28

6.2.1 Empirical determination of computational factors (e.g. singular values) 28

6.2.2 Computational time for dynamic corpora 28

6.2.3 Reasoning and logics 29

7 OTHER APPLICATIONS OF LSA 29

7.1.1 Community formation, community support and collaboration 29

7.1.2 Human Resource Management and task allocation 30

7.1.3 Localizing resources 31

7.1.4 Support of assessment and feedback 31

8 REFERENCES 31

9 GENERAL BIBLIOGRAPHY ON LSA 37

1 Introduction

This State of the art on Latent Semantic Analysis (LSA) captures current knowledge on and applications of LSA. Furthermore, it tries to connect this knowledge to other applications in education by identifying useful ways in which LSA can be used and what benefits it offers. Rather than being exhaustive, the deliverable tries to capture the essence of each topic. When appropriate, references for further reading are given.

After a short introduction to LSA, section 3.1 discusses Latent Semantic Indexing (LSI) as a first application of LSA. Further, several educational applications of LSA like intelligent tutoring systems, question answering and accreditation of prior learning are discussed. Chapter 4 will go into issues in text corpus construction such as corpus and document size, document selection, stemming and stopping strategies. Several methodological considerations are discussed in chapter 5 and chapter 6 is concerned with the strengths and weaknesses of LSA. The document is concluded by discussing examples of other applications of LSA in education.

2 What is Latent Semantic Analysis?

2.1 Definition

Latent Semantic Analysis (LSA), also referred to as Latent Semantic Indexing, is a technique for document retrieval, or more general document comparison, that is based on text vector representation. Text vector representations are based on representation of a corpus of text in a matrix of terms by documents with the cells in the matrix containing frequency measures for the terms (see Table 1).

Terms	Documents																
	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15	B16	B17
algorithms	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0
application	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
delay	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
differential	0	0	0	1	0	0	0	1	0	1	1	1	1	1	1	0	0
equations	1	1	0	1	0	0	0	1	0	1	1	1	1	1	1	0	0
implementation	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
integral	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
introduction	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
methods	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
nonlinear	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
ordinary	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0
oscillation	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
partial	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
problem	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0
systems	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0
theory	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	1

Table 1. Example of a term-document matrix (Berry, Dumais & O'Brien, (1994, p.8))

A document is then represented as a vector of term frequencies. Figure 1. depicts a two-dimensional plot of the terms and documents from Table 1. A vector is a line from the origin through a point representing a specific term or document.

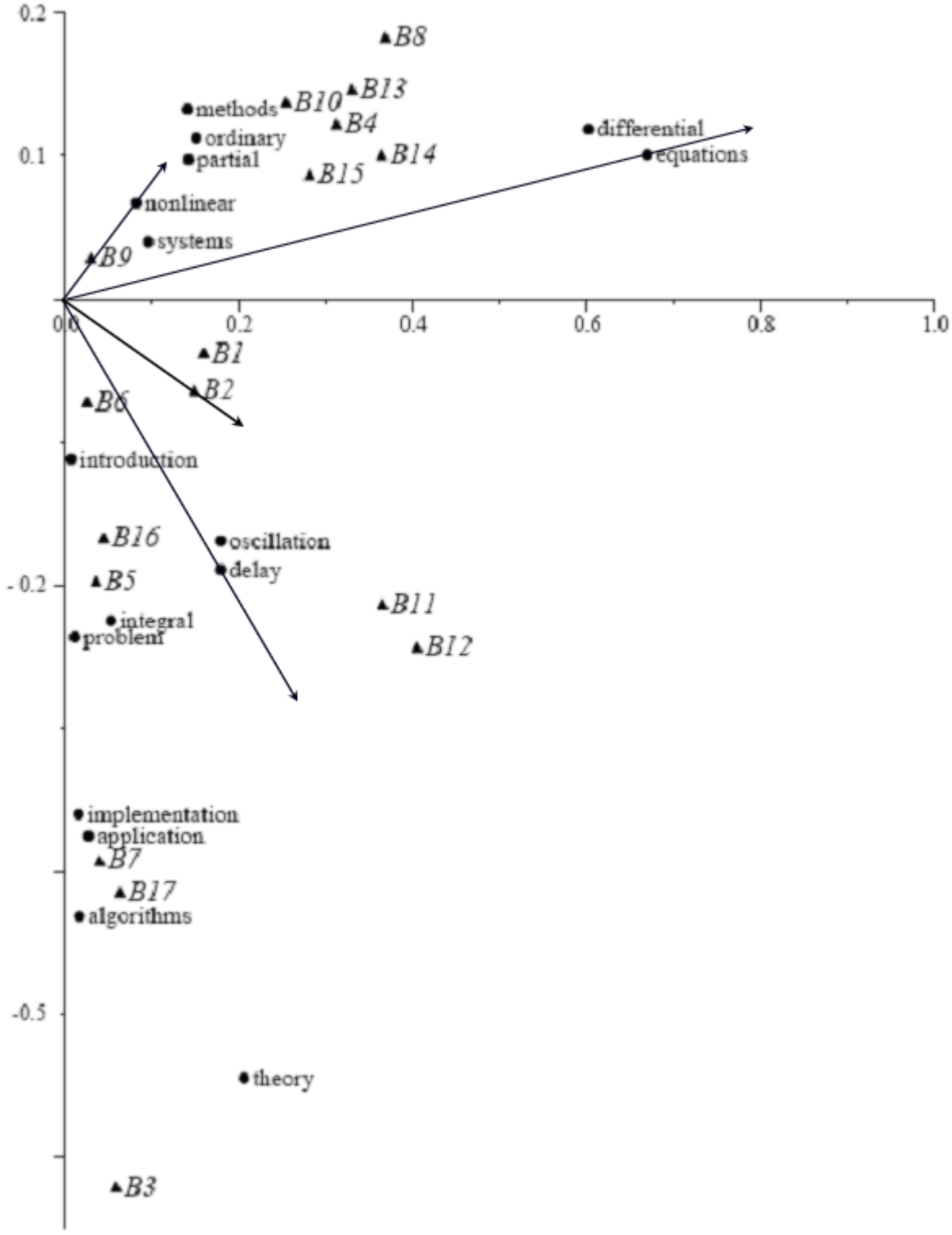

Figure 1. Two-dimensional plot of the terms and documents in Table 1. (adapted from Berry et al. (1994, p. 9))

Note that in this vector representation all syntactical information is lost (it is a 'bag of words' representation). Negation or qualifying information ("not true", "partially true") is not represented. As with any vector representation of documents, one can compute similarities between documents by computing the distance or the angle between their vectors. Latent Semantic

Analysis goes beyond these techniques in that it projects document vectors in a multidimensional space that is *abstracted* from the data.

LSA is a three steps procedure in which a data matrix is reconstructed using less dimensions than are present in the original data. In the first step, a dimensional model is obtained by performing a singular value decomposition (svd) of the data matrix. This returns a diagonal matrix with the singular values and two orthogonal rotation matrices. The number of singular values > 0 are the number of dimensions in the data. In the second step the number of dimensions is reduced by dropping from further calculations the smallest singular values and the corresponding rows and columns in the rotation matrices. In the third step the data matrices reproduced on the basis of the reduced model (Deerwester, Dumais, Furnas, Landauer & Harshman,1990). The underlying techniques as well as the interpretation of LSA bear close resemblance to Principal Components Analysis and Factor Analysis, both techniques that are more commonly used in social and behavioural sciences.

Semantically speaking, latent semantic analysis expresses the meaning of a text passage as a weighted sum of underlying constructs, such as contexts and concepts (Quesada, 2003). In this view, LSA is a technique to reveal these *latent semantic* variables and helps to explain the common core behind documents (context) and terms (concepts) (Landauer & Dumais, 1997). The extraction of the latent variables can be seen as a form of learning from observations. Figure 2 provides a schematic overview of how LSA is used to compare a query with a corpus.

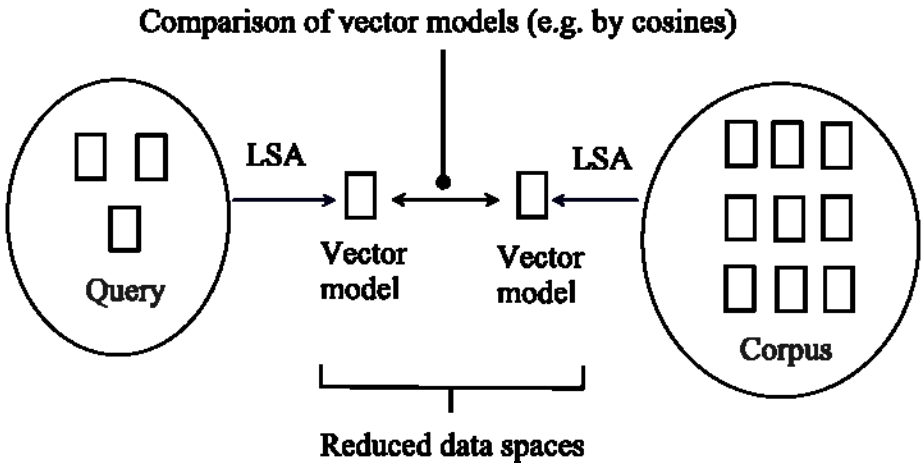

Figure 2. Schematic overview of the LSA process

2.2 LSA in a nutshell

LSA starts with a collection of terms and documents. The frequencies with which the terms occur in the documents are recorded in a table, the Term-Document matrix. A document is represented by a column vector of term frequencies (a document vector) and a term is represented by a row vector of frequencies across documents (a term vector). Note that the order of the concepts in the document is irrelevant: LSA does not log syntactic information. The dimensions of this Term-Document matrix, let's call it **T**, are reduced in two stages. In the first stage a singular value decomposition (SVD) of the data matrix is obtained. SVD can be seen as a generalization of principal components analysis (PCA) or factor analysis. In PCA a symmetrical matrix (e.g. a covariance matrix) is decomposed as **C** = **U Λ U'**, where **U** are the matrices with eigenvectors and **Λ** is the diagonal matrix with eigenvalues. In singular value decomposition a non-symmetrical, non-singular matrix **T** is decomposed as **T** = **L S R'** where **L** and **R** are orthonormal matrices and **S** is a diagonal matrix with singular values. The number of singular values > 0 is equal to the dimensions of the matrix. Think of the values of **S** as defining orthogonal axes in a high-dimensional space with the values corresponding to the length of the axes. To reduce the number of dimensions, only the longest axes are retained by removing rows and columns in **S** and the corresponding ones in **L** and **R'**. The original matrix is now reconstructed from these reduced matrices. In the reconstructed matrix, a document-vector may contain a frequency for a word *W* that did not appear in the original document. In other words, a query for "all documents about *W*" may return documents that do not contain the word W itself, but words that tend to co-occur with *W*. Several other measures can be obtained using the reconstructed Term-Document matrix, such as the correlation between document vectors. The higher the correlation, the more the documents resemble one another. That makes it possible to compare documents to each other, or compare a document to a vector of search terms.

Table 2 contains a fictitious example of six documents dealing with different kinds of apes and monkeys. Obviously, a real term-document matrix will contains far more terms as well as documents. The key term for the example is 'ape'. Although various documents deal with big apes", note that document 2 does not contain the term itself, although it refers several times to species of big apes.

Original term document matrix

doc nr.	1	2	3	4	5	6
gibbon	5	1	4	9	5	0
chimpanzee	4	4	2	0	2	3
orangutan	3	5	3	0	2	4
gorilla	4	4	2	0	2	3
ape	3	0	1	0	4	2
bonobo	4	6	2	0	2	2
mandrill	1	0	5	3	2	0
baboon	2	0	5	3	2	0
capuchins	0	0	0	0	6	0
douroucoulis	1	0	3	3	2	0

Table 2: term frequencies

A singular value decomposition returns the singular values presented in Table 3.

18.91337	0	0	0	0	0
0	10.95706	0	0	0	0
0	0	6.158899	0	0	0
0	0	0	4.252632	0	0
0	0	0	0	2.627552	0
0	0	0	0	0	1.518679

Table 3: Singular values of the term-document matrix

A crucial decision is the selection of dimensions to be retained in the further analyses. Several heuristics are suggested in the literature. Here we only demonstrate one that relies on visual inspection of the data: the Scree test (Cattell, 1966), which is also being used to determine the number of factors in a factor analysis. In figure 1 the singular values are plotted. Note the sharp decline after the fourth singular value. Based on this Scree-test, four singular values are retained.

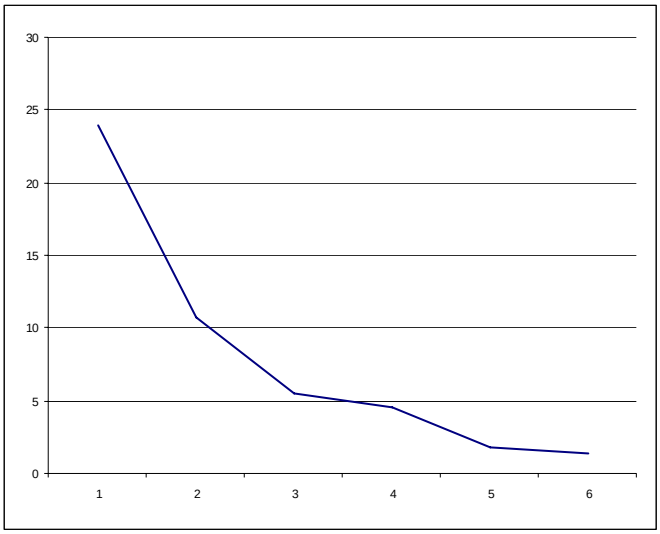

Figure 2: plot of the singular values

The dropped singular values as well as the columns and rows of the left and right matrices that correspond to the dropped values are greyed out in the tables.

-0.55021	-0.52272	-0.04811	0.611989	-0.0017	0.136955
-0.3159	0.328689	-0.02428	0.038847	0.19887	-0.06333
-0.34558	0.385562	-0.1094	-0.203	-0.08707	0.715652
-0.3159	0.328689	-0.02428	0.038847	0.19887	-0.06333
-0.22463	0.054634	0.467587	-0.07905	0.673995	-0.14646
-0.34191	0.398948	-0.1021	0.155244	-0.4806	-0.53049
-0.25714	-0.27118	-0.17856	-0.5379	-0.10488	-0.00798
-0.28339	-0.25524	-0.18854	-0.45367	0.102091	-0.36575
-0.1452	-0.07397	0.827694	-0.15621	-0.44535	0.04102
-0.20829	-0.23608	-0.05448	-0.17703	-0.09119	0.150591

Table 4: left matrix with dropped columns greyed out

-0.49641	-0.36253	-0.46196	-0.3806	-0.45772	-0.23321
0.174734	0.586681	-0.19232	-0.63812	-0.13508	0.403534
-0.0615	-0.22763	-0.3821	-0.27565	0.849614	0.023977
0.358195	0.197348	-0.76733	0.470791	-0.11071	-0.1003
0.543832	-0.65828	-0.01799	-0.11314	-0.19503	0.468783
-0.54333	0.016883	-0.12041	0.370833	0.010383	0.743224

Table 5: right matrix with dropped rows greyed out

Table 6 shows the reproduced term-document matrix. Note that ‘ape’ now has a reconstructed frequency of 1.17 in document 2, where the term does not occur at all.

	1	2	3	4	5	6
gibbon	5.12	0.99	4.02	8.92	5.00	-0.15
chimpanzee	3.66	4.35	2.00	0.09	2.10	2.83
orangutan	3.71	4.83	3.13	-0.43	1.94	3.30
gorilla	3.66	4.35	2.00	0.09	2.10	2.83
ape	1.92	1.17	1.01	0.28	4.35	1.34
bonobo	4.25	5.18	1.88	0.16	1.76	3.19
mandrill	1.14	-0.18	4.99	2.97	1.95	0.14
baboon	1.55	0.19	4.94	3.24	2.06	0.29
capuchins	0.67	-0.77	-0.01	-0.16	5.77	0.50
douroucoulis	1.25	-0.16	3.02	2.89	1.95	-0.06

Table 6: reproduced term-document matrix

LSA is sometimes presented as a statistical technique, but this is slightly misleading. LSA is primarily a mathematical technique. However, the core of the method, singular value decomposition, is a least squares technique that assumes, or at least performs best when the data is normally distributed and one may question whether such is the case with term frequencies (Rosario, 2000). LSA can be applied under different assumptions regarding the underlying data and this probabilistic approach may yield better results that are interpretable in a statistical way (Hofmann, 1999). Since, however, the applications of LSA reviewed have nearly all used the classical approach to LSA, we will not elaborate probabilistic approaches any further.

When we say that LSA allows comparisons of documents, the term documents should be considered in a broad sense, to cover text ranging from utterances, including search queries and sentences, to complete books. Several application areas of LSA stem from this basic approach. Thus, LSA is being used to query text databases (Berry, Dumais & O'Brien, 1994; Giles, Wo & Berry, 2001), to determine coherence within text passages or between chapters in a book (Foltz, Kintsch & Landauer, 1998), to grade essays after comparing them to one or more standards (Foltz, Laham & Landauer, 1999), to select (Wolfe et al., 1998) and sequence learning material (Zampa & Lemaire, 2002), to compare and match task and job descriptions and workers (Laham, Bennett & Landauer, 2000) or to use LSA based comparison of learning material as a basis for accreditation of prior learning (van Bruggen et al., 2004).

The application areas for latent semantic analysis can be grouped into *information retrieval* (where it is generally called latent semantic indexing), *cognitive science* and *education*. In our review we will concentrate on the first and third application areas. Uses in cognitive science, such as language learning, representation of semantics as well as problem solving and concept learning, are largely ignored in this report (see 0 for discourse processing however).

Next to the application areas, we discuss implementation issues, in particular topics around corpus construction, and some methodological considerations. Core to the latter is the determination of the numbers of dimensions to be used in the reproduction of the data.

3 Application areas of LSA

In the past ten to fifteen years the LSA technique has been applied in a wide range of domains. In this chapter we will describe these domains more or less in chronological time order of these implementations.

3.1 Document retrieval and latent semantic indexing (LSI)

The technique of Latent Semantic Analysis (LSA) was initially applied in the field of document retrieval (Deerwester, Dumais, Furnas, Landauer & Harshman, 1990; Dumais, 1992; Dumais, 1997; Berry, Drmac & Jessup, 1999; Letsche, 1997; Rosario, 2000; Chung-Min, Stoffel, Post, Bassu & Behrens, 2001; Freeman, Thompson & Cohen, 2000).

Deerwester et al. (1990) proposed a new approach to automatic indexing and information retrieval on the internet in the 1990's. They describe the search problem at that time as:

> users want to retrieve on the basis of conceptual content, and individual words provide unreliable evidence about the conceptual topic or meaning of a document. There are usually many ways to express a given concept, so the literal terms in a user's query may not match those of a relevant document. In addition, most words have multiple meanings, so terms in a user's query will literally match terms in documents that are not of interest to the user (p. 1).

This statement requires some explanation on how most search engines work.

Latent semantic indexing adds an important step to the document indexing process (Deerwester et al., 1990). In addition to recording which keywords a document contains, the method examines the document collection as a whole, to see which other documents contain some of those same words. It is assumed that there is some underlying, latent, semantic structure in the data that is partially obscured by the randomness of word choice with respect to retrieval. Mathematical techniques are used to estimate this latent structure, and get rid of the obscuring "noise" (e.g. common words like 'the', 'a', 'an'). Retrieval occurs by projecting the query vector (which can range from keywords to documents) on the latent structure and calculating the angle between the query vector and the document vectors, which form the latent structure. The document vectors with the smallest angles are returned. Because two documents may be semantically very close even if they do not share a particular keyword, LSI does not require an exact match to return useful results. Yu et al. (2005) state that "*where a plain keyword search will fail if there is no exact match, LSI will often return relevant documents that don't contain the keyword at all" (p. 7)*. Instead it returns documents that have overlapping (context) vectors with the used keyword. LSA will perform better if more keywords are used, thus providing context information for the search and mapping more document vectors.

Using LSA for indexing can significantly improve three important characteristics of a search engine (Yu et al., 2005): *recall* (find every document relevant to the query), *precision* (no irrelevant documents in the result set) and *ranking* (most relevant results come first).

3.1.1 Synonymy and polysemy

Synonymy and polysemy are two important issues in retrieval methods. Synonymy means that many different words can refer to the same concept.

Deerwester et al.(1990) state that "*a fundamental deficiency of many information retrieval methods is that the words searchers use often are not the same as those by which the information they seek has been indexed" (p. 1).* They also found that "*the degree of variability in descriptive term usage is much greater than is commonly suspected. For example, two people choose the same main key word for a single well-known object less than 20% of the time*"(p. 1). They also mention studies that reported similar poor inter-rater agreement consistency and inconsistency in the generation of search terms, especially by expert intermediaries or less experienced searchers. They conclude that "*the prevalence of synonyms tends to decrease the "recall" performance of retrieval systems*" (p. 1).

The phenomenon of polysemy means that most words have more than one distinct meaning (for example the words "arm" or "safe"). Deerwester et.al. (1990) state that "*The use of a term in a search query does not necessarily mean that a document containing or labelled by the same term is of interest. Polysemy is one factor underlying poor "precision" of search engines*" (p. 2).

Deerwester et.al. (1990) identify the polysemy problem in information retrieval as a problem which can only be dealt partially with by the LSI method. They state that:

> While the LSI method deals nicely with the synonymy problem, it offers a partial solution to the polysemy problem (Deerwester et al., 1990b), since the meaning of a word is determined not only by other words in the document but by other appropriate words in the query not used by the author of a particular relevant document, i.e. there is context-dependency when it comes to document retrieval. The term itself however is represented as a single term vector in the space. That is, a term with several different meanings (e.g. "bank"), is represented as a weighted average of the different meanings. Kintsch' (2001) work on 'predication' combines the relations between terms found using LSA with context-dependent information that through a spreading activation model strengthens some of the relations, while inhibiting others. Hofmann (1999) claims that 'probabilistic LSA', which is based on a latent class model (rather than the continuous model of SVD) and multinomial distributions of term occurrences was able to make a distinction between the different meanings of polysemic words *(*p. 21).

Deerwester et.al. (1990): *The latent semantic indexing methods* […] *are capable of improving the way in which is dealt with the problem of multiple terms referring to the same object* (p. 22) (synonymy) while performing searches. This is further described as:

> [the LSI methods] replace individual terms as the descriptors of documents by independent "artificial concepts" that can be specified by any one of several terms (or documents) or combinations thereof. In this way relevant documents that do not contain the terms of the query, or whose contained terms are qualified by other terms in the query or document but not both, can be properly characterized and identified. The method yields a retrieval scheme in which documents are ordered continuously by similarity to the query, so that a threshold can be set depending on the desires and resources of the user and service" (p.22).

3.2 Representation of semantics and discourse processing

Foltz (1996) and Landauer & Dumais (1997) extended the LSA technique to discourse analysis and problems with learning and language processes. They were mainly concerned with questions as: How do humans derive meaning from texts? What factors influence a reader's ability to extract and retain information from textual material? The fact that a LSA-model could 'understand' meaning of text without word order is one of the main fascinations these researchers had. They used LSA to extract and represent the contextual-usage meaning of

words. The underlying idea is that the aggregate of all the word contexts in which a word does and does not appear provides a set of constraints that determines the similarity of meaning of words and sets of words to each other.

The adequacy of LSA's modelling of human knowledge has been established in a variety of ways (Landauer, 2002b). Landauer & Dumais (1997) compared human judgments of student essays to measures derived from a LSA-model, in this way providing evidence that information about the meaning of passages (semantics) may be carried by words independently of their order (syntax). They found that LSA-based measures- which take no account of word order – were as closely related to human judgments as the human judgments were to each other. They also found that LSA measures predicted external measures of the same knowledge as well or better that the human judgments'. They experimented further and found that the vectors for words derived from an encyclopaedia analysis predicted the correct answers to standardized vocabulary tests in which students are asked to judge similarity of meaning. LSA simulations matched the performance of moderately competent students. They also demonstrated that LSA 'learned' word meanings perform reading at about the same rate as late primary school children.

Foltz (1996) did something similar as the 'students essay experiment' of Landauer & Dumais (1997). He used LSA similarity measurements (in two of the three experiments by using cosines') to analyse students' essays to determine what a student learned from the original text, which texts influenced their summaries (thus predicting the source of students' knowledge) and for grading the quality of information cited in the essay. The grading was done by comparing the semantic overlap between the students essay and the original source text and between the student essay and the 10 sentences an expert grader thought were most important. A grade was assigned to each essay on the basis of the mean of the cosines between each sentence in the essay and the closest of the 10 sentences chosen by the expert grader. Next to this, in the third experiment, Foltz (1996) also used LSA to measure the coherence and comprehensibility of texts. Foltz (1996) argues that the "*coherence of text can be calculated by examining the repetition of referents used in propositions through the text*" (p.8) and that the "*degree of repetition of arguments in a text is highly predictive of the reader's recall*"(p.8.), thus improving quality of texts. Foltz' (1996) LSA predictions on the coherence of text were made by:

> […] calculating the amount of semantic overlap between adjoining sentences in the text. Thus, for each text, the cosine distance was computed between the vector of sentence N and the vector for the sentence N+1. The mean of all the cosines for a text was then calculated to generate a single number representing the mean coherence for a text (p. 8).

Foltz (1996) concludes that text coherence could be measured on a local (coherence between sentences) and a global level (overall coherence of the text), only that the grain size for prediction is larger for LSA than for a proposition method determining coherence of text. He states that:

> […] propositions represent semantic information at a clause level, while LSA is more successful in performing analyses at a sentence or paragraph level. The few words in a clause make the vectors in LSA highly dependent on the words used in that clause, whereas sentences contain enough words to permit a vector that more accurately captures the semantics of the sentence" (p.9).

This application makes it possible to easily detect incoherence of text and repair them, thus improving readability and learnability of texts.

Landauer & Dumais (1997) stated that the fact that LSA can capture as much of meaning as it does without using word order shows that the mere combination of words in passages constrains overall meaning very strongly. They also state that this effect depends on the dimensionality of the representation (also see Landauer, 2002b). Nonetheless they underscore

the importance of syntax for the human meaning making processes, hypothesizing that it may reduce working memory load or ease the construction of sequential utterances.

Walter Kintsch (1998) argues that situational as well as semantic contexts influence the meaning of a concept. He states that "*concepts and their meaning expressed in propositional representations had been hand-coded until then, hence it was difficult to use them in large-scale, practical applications*" (p. 417). He proposed LSA as an alternative to make these propositional representations, thus representing the concept in the derived LSA-vector (node in the propositional network) and the neighbouring vectors in that space as the context of a concept.

Laham (1997) also compares LSA vector space with human knowledge organization and he also argues that LSA could help to organize related concepts, by clustering concepts, which have a small difference in cosines' between the LSA-vectors. This seems to work better for certain categories of words than for others (e.g. comparison of nature related words to man-made artefacts). After discovering LSA vector space as a possible representation of human knowledge organization, research was focused further on four different and more specific semantic problems: metaphor interpretation, causal inferences, similarity judgments and homonym disambiguation (words spelled and pronounced alike but different in meaning, e.g. cleave meaning "to cut" and cleave meaning "to adhere"). On the areas of metaphor interpretation, similarity judgments and homonym disambiguation's advances are made in the past years. Kintsch (2001) has developed an extended model called the 'predication model' to improve the performance of LSA compared with human performances on language comprehension. He computed the meaning of sentences with LSA, but in a context-related manner: he adjusted word vectors contextually according to their syntax in a sentence and then summed them up to compute a sentence vector. Sentence vectors of the form N1-is-N2 were computed by modifying the predicate vector N2 according to the argument vector N1. Thus, a context appropriate sense of the predicate is generated. In this way he was able to improve the LSA performance on metaphor interpretation, reaching almost similar judgments and patterns compared to humans (Kintsch & Bowles, 2002). The LSA-model also had problems with difficult metaphors, like humans, but was able to solve more easy metaphors and reaching logical solutions for the more difficult ones.

The problem of causal inferences remains a recurrent problem with the use of LSA, because it does not reckon with syntactic order and relational propositions. Kanejiya, Kumar & Prasad (2003) have been suggesting an extended model as well to solve this problem, called Syntactically Enhanced LSA (SELSA). Their approach generalizes LSA by

> […] considering a word along with its syntactic neighbourhood given by the part-of-speech tag of its preceding word, as a unit of knowledge representation" […] It also provides better discrimination of syntactic-semantic knowledge representation than LSA, but has not yet been highly successful in experimental setting. In an experiment with Auto-tutor SELSA was able to correctly evaluate a few more answers than LSA but is having less correlation with human evaluators than LSA has.*" (p.1*).

Future research is needed in this area.

3.3 Educational applications

From its original inception in information retrieval, LSA has found wide application in research areas as cognitive models of human word meaning acquisition (Landauer & Dumais, 1997) and language understanding (Kintsch, 1998; Wiemer-Hastings & Zipitria, 2001), as described in the previous paragraph. Here we review applications of LSA in educational settings. Stahl (1997, in Lemaire & Dessus, 2001) suggests that LSA may be appropriate in several ways. The first concerns automatically assessing essays and providing feedback to students. The aims of the

systems developed here may vary from providing (support for rating) a summative evaluation to offering formative support to students who are preparing essays or summaries. The second type of LSA educational application involves modelling the knowledge of the learner in order to select and sequence suitable instructional materials. Here, LSA is used to model both learners and instructional materials in the same multidimensional semantic space, making it possible to assess similarities between the two. The key challenge in this type of application is to select material that is in the "zone of proximal development", thus providing the student with the right amount of new information. The third type of application is to use LSA to connect students with each other and with relevant experts, thus facilitating community formation and question answering. LSA can assess the areas of interest or the level of knowledge from the users based on their products and then suggest matches. Recently, other ideas for applications have emerged in the field, such as possible usability for accreditation of prior learning.

3.3.1 Intelligent tutoring systems- assessment and feedback of free text responses

LSA has been used to assist in various assessments with various aims: helping tutors to assess students' performances, but also helping students to reach an optimal performance by providing feedback while they were practicing. Both applications are described in this paragraph. Although a difference between assessment and feedback is made, many applications can be used in both ways. The line to draw the difference between 'assessment' and 'feedback' considering these applications is very thin. Miller (2003) reviews contemporary essay-scoring systems built on LSA and mentions the Intelligent Essay Assessor, Summary Street, State the Essence, Apex and Select-a-Kibitzer in one breath. The difference made below is one mainly based on initial orientation within these projects and their (initial) main research focus in empirical experiments.

Assessment

LSA has been used to grade essays. Foltz (1996) compared essay scores assigned by humans to those assigned by LSA and found little difference. Foltz concluded that, at least, "*LSA is an automatic and fast method that permits quick measurements of the semantic similarity between pieces of textual information*"(p.10), thus allowing it to be used as a means to grade essays by correlating text similarity to essays of known quality. Landauer, Foltz & Laham (1998), following up on this approach, describe several approaches to automatic essay evaluation. LSA can compare the essay with defined standard(s), e.g. written by expert writers, written by previous students with a high grade, that is the so called 'golden standard' approach. LSA would then grade a student's essay by applying a function on the cosines between the essay and one or more standards. Grading performance was improved by combining, with roughly equal weights, the cosine measure and a vector length measure: the former measure is sensitive of the content of the essay, the latter to the amount of content (Kintsch, 2002a). Landauer, Foltz & Laham, 1998) compared human and machine ratings using different scoring schemes such as holistic scores and topic scores. Eventually, this research led to the LSA application for automatic assessment that is probably best known in the educational community, the **Intelligent Essay Assessor** (IEA) (Foltz et al., 1999). As with any LSA application, the Intelligent Essay Assessor is trained on material drawn from the domain of the essay topic. IEA does not require a large set of graded essays. Tuning the system may require just a few examples, including a so-called "golden standard". IEA has been found to rate essays with a reliability that matches those of human raters (Foltz, Gilliam & Kendall, 2000).

Another application aimed at assessing student performances in essay writing is **Apex** (Assistant for Preparing Exams) (Dessus, Lemaire & Vernier, 2000; Lemaire & Dessus, 2001), developed at the Université Pierre Mendés France in Grenoble (Miller, 2003). Miller (2003)

states that Apex "*uses LSA to assess student essays on topic coverage, discourse structure and coherence*"(p.22). According to Miller (2003) Apex differs from Intelligent Essay Assessor and Summary Street (see feedback):

> […] in that the partition of the source text in topics is much more fine grained. For calibration, the teacher must identify notions – short passages of text which exposit a certain key concept – and the topic or topics to which each notion belongs. For an essay on a given topic, Apex computes the cosine coefficient between the essay and each relevant notion (p.23).

and the average of these cosines' form the final score. Apex scores' were found to correlate well with human scores for content and overall essay quality. By using the notions, Apex is able to construct an overview of the structure of an essay, thus helping students in planning the discourse and also highlighting areas of concern. Miller (2003) states that:

> [this] outline is produced by having LSA find and print each essay paragraph's closest corresponding notion. If no notion correlates above a certain threshold, the paragraph is flagged as potentially irrelevant. The completed outline also helps to identify repetitious sections. Apex also performs coherence analysis (comparative to (Foltz et al., 1998) by comparing adjacent sentence pairs and reporting abrupt topic shifts) (p.23).

Feedback

Several projects attempted to use LSA to provide (faster) feedback to support self study. The feedback allows students to engage in extensive independent practice without placing excessive demands on teachers for feedback. Detailed feedback often requires that LSA operates on more fine-grained aspects of texts. For instance, the coherence of a text has been measured by calculating cosine similarities between individual sentences. A high overall similarity indicates repetition or rephrasing of the text, while an overall low similarity is an indication that the text has a low coherence. Drops in similarities between successive sentences can indicate topic breaks. A high average number of topic breaks may indicate that a text jumps from topic to topic.

These types of measures are used in **Select-a-Kibitzer** (Wiemer-Hastings, Wiemer-Hastings & Graesser, 1999; Wiemer-Hastings & Graesser, 2000), an educational software system that provides feedback on student compositions. Once a student has entered her text, specialist agents - the Kibitzers - may be invoked to provide feedback on the particular text characteristics in which each of them specializes. Each kibitzer acts as a critic for a particular discourse feature, be it stylistic, grammatical or semantic. LSA is used to determine the coherence of the text, and the topic breaks between the sentences are used to identify semantic chunks. The sentence with the highest average similarity to other sentences in the chunk is considered the key sentence and is presented as the system's understanding of the topic. Miller (2003) states that:

> like Apex, Select-a-Kibitzer generates outlines of essays, but it does so without reference to a source text. The program uses clustering methods on the LSA semantic space (like Laham, 1997) to identify discrete topical chunks in the corpus. For each chunk, the program selects as an archetypical sentence the one that compares best with all other sentences in the chunk. An outline of key points is then produced by printing the selected sentences in order of appearance in the essay. This outline gives the student an idea of the essay's progression of ideas, something particularly useful for beginning writers (p.24).

The LSA engine in Select-a-Kibitzer is also trained in template sentences to help determine the purpose of sentences. Templates such as "I would change….because" are used to indicate why-reasoning.

Miller (2003) also describes **State the Essence** (Kintsch et al., 2000) and states that it "*was designed to improve elementary schools student's summarization skills, helping them mediate the conflict between concision and comprehensiveness). After an initial spell-check, LSA is used to measure topic coverage, irrelevancy and redundancy* (p.21). This system does not provide feedback on other aspects of writing, such as sentence structure, organization and style. Because of LSA's ignorance of syntax or morphology (Miller, 2003), it cannot judge most matters of mechanics and style (e.g. spelling, grammar, clichés, tense shifts). Miller (2003) states that:

> […] students can revise and resubmit as often as they like. Once they are satisfied with the feedback, they submit their papers to the teacher for complete grading. Initial trials of State the Essence indicated areas for improvement. Overall correlation with humans was inconsistent and there was no evidence that use of the program resulted in increased writing skills or learning. More seriously, students tended to forget or ignore the fact that the program was evaluating content only, preoccupation with the numerical score incited many students to abandon good writing style in favour of increasing their score by the cheapest means possible. They received heavy penalties for organization and mechanics upon human grading *(*p.21*).*

Miller (2003) mentions that "hypothesizing that the bulk of the problem lay in the feedback mechanism, Kintsch et al. (2000) revised the system"(p.22). Visualizations of numeric scores and changes on how and when advice on redundancy is presented were made. This new version was renamed to Summary street.

Summary Street provides various kinds of immediate feedback, primarily about whether a student summary adequately covers important source content (based on several 'golden summaries' and representations of the source text) and fulfils other requirements, such as length. It tells students what information in the source is missing, provides comments on redundancy and relevance.

Experiments with Summary Street suggest that it is especially helpful when students are faced with more difficult tasks or with a harder text. Kintsch (2002b) and Wade-Stein & Kintsch (2003) reported three notable results of using system feedback while writing summaries. Time on task increased significantly when students could use the system: students were willing to work harder and longer when given immediate feedback. Summaries written with content feedback received higher grades form the teachers. This was the case for difficult summaries, for which grades more than doubled, whereas for texts that were easy to summarize, the use of the system had no significant effect. The researchers also observed a transfer effect. Students who had written a summary in the previous week with the help of *Summary Street* wrote better summaries even when they no longer had access to the feedback the system provides.

Magliano, Wiemer-Hastings, Millis, Munoz & McNamara (2002) tested a computer-based procedure for assessing reader strategies that was based on latent semantic analysis (LSA). During a computerized version of self-explanation-reading-training (SERT), students read texts and typed self-explanations. Self-explanation refers to explaining difficult text to oneself. The strategies include using logic or world knowledge to elaborate on the current sentence (knowledge building explanation), making conceptual bridges among ideas in text and predicting what will come next in the text (sentence-focused explanation). A minimalist approach would be to paraphrase the sentence or provide a vague description. The goal was to test if LSA could be used to assess the extent to which students used these strategies and classify the self-explanations as 'knowledge building', 'sentence focused' or 'minimalist'. Several semantic benchmarks were used as a reference model for students' self-explanations: (1) the current sentence (2) causally important prior sentences (3) relevant world knowledge and sources that the reader can draw upon while self-explaining. For the last, a semantic space on heart diseases created by researchers at the University of Colorado at Boulder was used (also see http://lsa.colorado.edu). The hypotheses were that knowledge-building self-explanations should have a high overlap (e.g. high cosines) with causally important information from the prior

text and/or relevant world knowledge. In contrast, a minimalist self-explanation should have a relatively low overlap with the prior text and relevant world knowledge, but a relatively high overlap with the current sentence, because the reader is primarily paraphrasing the sentence. A sentence-focused self explanation should also have a relative high overlap with the current sentence, but should have intermediate overlap with the prior text and relevant world knowledge. The LSA-assessment was compared with human judgments of the self-explanations. Magliano, Wiemer-Hastings, Millis, Munoz & McNamara (2002) state that:

> […] both human judgments and LSA were remarkably similar and indicated that students who were not complying with SERT tended to paraphrase the text sentences, whereas students who were complying with SERT tended to explain the sentences in terms of what they knew about the world and of information provided in the prior text context *"* (p.181).

3.3.2 Intelligent tutoring systems- selection and sequencing of instruction

LSA has been used to select instructional text that is appropriate to the student's background knowledge, i.e. a text that matches the prior knowledge of a student partly, but also adds new concepts to it. Appropriate text is neither too easy, nor too hard for a student.

Rehder et al. (1998), Wolfe et al. (1998) and Landauer (2002b) began to use LSA to match students with text at the optimal level of conceptual complexity for learning. LSA was used to characterize both knowledge of an individual student before and after reading a particular text and the knowledge conveyed by that text. Wolfe et al. (1998) addressed a similar issue, referring to it as the "zone-of-learnability". The key to their approach was to select a study text to match the prior knowledge of the learner as closely as possible. First, they collected data on the students' prior knowledge, and then had the students study one of four different texts about, a topic such as the anatomy, function and purpose of the human heart and the circulatory system. The texts ranged in difficulty from elementary school to medical school level. As expected, learning gains were related to prior knowledge: texts that were too easy or too complex yielded weaker learning gains. Wolfe et al. (1998) presented a number of curve-fitting solutions that relate LSA-based similarity measures between prior knowledge and the study texts to predict learning effects. If the cosine between an essay written by a student and an instructional text was moderate, learning was successful (around 40% improvement in test scores or essay grades between pre- and post-test); when the cosine was too low (not enough background knowledge) learning was poor, when the cosine was too high (not enough new information in the text), learning was equally poor. Zampa & Lemaire (2002) used LSA in an intelligent tutoring system to model a domain and the student and select appropriate texts to match students knowledge level. In their model, a domain is built of "lexemes", being either words in a language-learning domain, or facts and conclusions in a problem-solving domain. Note that this domain representation is not based on raw text, but requires prior identification of the lexemes. The student, it is assumed, learns the domain by being exposed to a series of lexemes. The tutoring system selects those texts/topics in a zone around the student and domain sequences that have already been addressed. Sequences that are too close or too remote are expected to yield a weaker learning effect and are therefore ignored.

AutoTutor (Wiemer-Hastings, 1999; Wiemer-Hastings et al., 1999) engages students in a natural language conversation and thus encourages them to provide elaborate answers to the questions it poses. AutoTutor scores the quality of the answers that the students provide in conversational turns using a variety of techniques, including LSA. AutoTutor rates the quality of the students' assertions much the same as intermediate-level experts, but not as well as accomplished experts. The LSA component of AutoTutor is able to discriminate between classes of simulated students, and is capable of tracking the increased coverage of a topic in successive turns. Lemaire & Dessus (2001) state that "*another purpose of Autotutor is to answer unrestricted student questions. It does so by selecting the closest piece of text to the question.*"(p.4). Recently (Graesser et al., 2005), implementations of AutoTutor in different

domains have been made and LSA has generally been successful in evaluating the quality of student explanations and assertions in tutorial dialog.

3.3.3 Community formation and community support

Stahl (1997) already mentioned possible usability of LSA to connect students with each other and with knowledge experts. (Yukawa, Kasahara, Kato & Kita, 2001) have implemented this idea in an expert recommendation system. They describe the system as that it:

> […] processes the description of a technical topic as input and then find engineers who have a high level of expertise in that area. The technique is an extended vector space model that locates both technical topics and engineers in the same multi-dimensional space and then calculated their relevance. This system can also retrieve engineers or documents that are related to a field matching a given engineer's technical interests (p. 1).

Recently this idea is elaborated upon for transient communities: communities that fulfil a specific goal and exist for a limited amount of time. Interest as well as expertise areas of members could be based on comparison of the document sets that members of the learning communities collect and produce (Kester et al. 2005, submitted). In this way, ad hoc, transient communities could be formed, based on questions asked by a community member and answered by an adhoc formed group of peers, who have knowledge on the topic and are stimulated by 'seeds' (little fragments of contents which seem to be relevant and are selected with the help of LSA). Kester et al. (2005) proposed a model for this and are planning to experiment with an implemented version of this model within the Agents for Support Activities (ASA) project (Croock et al., 2003) at the Open University of the Netherlands.

3.3.4 Question answering

Within the same ASA-project one model for an agent for support activities is based on question-answering (Croock et al. 2003). The basic idea is that students pose natural language questions to a database and the database will provide the most relevant (part of a) document to provide an answer. This is not a new idea. Caron (2000) reports on a prototype system for technical support called the Frequently Asked Question Organizer (FAQO). Caron (2000) states that "*this application enables technical support personnel to construct a knowledge base from email archives and other existing documents. Users can query the knowledge base using natural-language questions in order to find relevant documents"* (p.1). The prototype that used LSA for query matching outperformed the keyword-search tool that was previously used. In a recent experiment, community information is used as an additional source of information to specify context for a certain question (Almeida & Almeida, 2004). The community-based information was used in order to provide context for queries and influenced by recent interactions of the user with the service. The algorithm used was tested on the service of an online bookstore. The quality of content-based ranking strategies in this way could be improved significantly and retrieval was improved with 48%.

3.3.5 Accreditation of Prior Learning and positioning

Van Bruggen et al. (2004) and Koper, van Bruggen, Rusman & Giesbers (2005) propose to use LSA to position learners in learning networks. Because learners can enter and leave such a network as they like, there is a recurrent need to position them in the right place within the network. In order to prevent the learner from taking redundant or too complex learning material and to accredit prior learning, latent semantic analysis is proposed as a tool for learner positioning in learning networks (van Bruggen et al., 2004). The core assumption is that equivalence of outcomes will be reflected in, or can be approximated by, the similarity of the contents of (learning) materials studied or produced by the student (source material) and the material contained in the learning activities in the learning network (target). LSA is used to

compare the contents of a learner's portfolio with the contents of learning materials contained in learning activities.

3.4 Human Resource Management

Another domain in which LSA is applied is Human Resource Management. A prototypical tool (HEADHUNTER) that matches jobs, people and instruction is worth mentioning here. Laham et al. (2000) experimented with LSA to match jobs, people and instruction in an air-force setting. Their aim was "*to help identify required job knowledge, to determine which members of the workforce had required job knowledge, pinpoint needed content which could be (re-)used within training settings and to maximize training and retraining efficiency."*(p.171). They processed data on three Air Force occupations for which full job descriptions were available. They then analysed "duty lists", tasks grouped into functional units, and individual tasks along with the tasks, which were actually completed in practice, thereby constructing a single semantic space for jobs and people. The semantic similarities between jobs and people could be used to decide between candidates for the job or to select a replacement.

It appeared that LSA could help to characterize tasks, occupations and personnel and measure the overlap in content between instructional courses covering the full range of tasks performed in many different occupations, thus indicating where the wheel was invented twice in the same working and training setting. Laham, Bennett & Landauer (2000) showed that their method could estimate

> […] the similarity of each task or occupation to every other task or occupation, measure the degree of match of each airman to every task of occupation, estimate which airmen could most easily take the place of others and indicated that LSA has the potential to identify in detail possible re-usable knowledge components and match the knowledge components required by new systems with those contained in segments of existing training materials and with the experience of individual airmen (p.173).

The experiment was based on a database with 20000 documents. They also emphasize the importance of applying LSA on large databases (e.g. for thousands of personnel in large braches (e.g. military or international corporations). LSA allows analysis that have been heretofore impossible because of the size and complexity of the data involved. Laham et al. (2000) also suggest that the system could also be trained to predict which course would bring a person closer to a target job profile.

Two experiments with a later version of this agent software, called CareerMap, again in an Air Force setting, are reported in Laham, Bennett & Derr (2002). In the first experiment, LSA is used to analyse course content and materials that are used in the current training settings and to identify appropriate places in alternative training structures where that content can be reused. They state that "*this saves time for training developers since the pre-existing content has already been validated as a part of its earlier application."* (p.1). Also gaps in the content for the new training structure become readily apparent. The second experiment is an implementation of a combined speech-to-texts (verbal communications translated to text) and LSA-based intelligent software agent for "*embedding automatic, continuous and cumulative analysis of verbal interactions in individual and team operational environments.*"(p.1). Currently it is impossible to evaluate verbal communication to identify critical information and content required to operators. Laham, Bennett & Derr (2002) state that LSA has potential for

> […] assisting operators in the performance of their tasks because it can 'listen' and in almost real-time evaluate free-form communication from a variety of sources and match content to stored language dictionaries. One application of this technology being explored is tracking and scoring the tactical communications […] to identify areas of training need and as an additional tool for assessing the efficacy of scenarios and missions." (p.1.)

Both experiments reported positive results with the use of LSA.

4 Implementation issues: corpus construction

LSA requires a text corpus and this chapter discusses several topics to consider when creating a corpus, such as corpus size, document size and document selection and related issues such as filtering and tidying, including stemming and stopping.

We do not discuss attempts to include additional, semantic information in corpora. Some semantic pre-processing of the corpus by identifying lexemes (words) (Zampa et al. 2002), by segmenting and breaking down corpus elements (sentences) by hand (Wiemer-Hastings, 1999) or by using an (automated corpus training) method of speech-tagging (Wiemer-Hastings & Zipitria, 2001) is reported in the literature. The LSA performance of both did not match human judgments as close as standard LSA did. This type of pre-processing is therefore omitted in our further discussions.

4.1.1 Corpus size

Most discussions of corpus construction are concentrated on the size of the corpus. However, it is not always clear what is meant by “large” and “small”. The same is true for what size is perceived as a minimum and/or maximum requirement to successfully apply LSA.

LSA is often used for document retrieval from very large document databases, containing ten thousands of documents and an input of 5000 documents to train LSA on the domain is quite common in these applications. Deerwester et al. (1990) state that a “reasonable size” is 1000 to 2000 abstracts which means about 5000-7000 index terms. In contrast, in Laham et al. (2000) a total number of 20.000 objects is mentioned as a very small dataset compared to LSA’s capabilities, but enough to do a fair job in estimating statistical regularities. Many authors seem to take this as a minimum requirement.

Several researchers show that big corpora are better, *but*, according to Landauer et al. (1998) the *goal* of using LSA is an important factor. They would like to “*truly represent the sum of an adult’s language exposure”* (p.35) of which they state that it is impossible because (1) it’s impossible to gather such a big corpus and (2) current computational power is not enough to perform SVD on 100.000’s x 10.000.000’s matrices. Educational uses of LSA, in contrast, are often confined to smaller corpora, that are more specific to (sub)domains. Within these corpora LSA has been shown to be robust against decreasing the size of the corpus (Wiemer-Hastings et al. 2000). According to Wiemer-Hastings et al. (2000), the best corpus is specific enough to allow for subtle semantic distinctions within a domain, but is general enough to ensure moderate variations in terminology won’t be lost. They report a ‘graceful degradation’ of performance. When they reduced the size of their text corpus from 2.3 MB to a minimum of 15 %, the performance of LSA in terms of correspondence with human raters decreased 12%. These results are clear indications that LSA can perform reasonably well in small scale corpora. More empirical research is desirable, especially because: *“A smaller corpus takes less time to train, less storage space, and less processing time for comparisons. Thus, if there is no significant performance advantage with larger corpora, they can be avoided”* (Wiemer-Hastings & Graesser, 2000, p.7).

Further research confirms that it is possible to obtain meaningful LSA results from smaller corpora. For example, Wild et al. (2005) used 43 files each consisting of a students’ answer on a marketing question from a real world exam. They performed many test runs (2016 in total) to

see what the influence of different parameters like pre-processing on the correlation between machine scores and human scores would be. Results show significant correlations between these scores, which means that LSA can work well on small corpora. Our own experience shows that meaningful results can be obtained by using 287 documents (about 10000 terms), each consisting of a single paragraph of text on monkeys and/or apes.

4.1.2 Document selection

Document selection mainly concerns the question whether it is better to have a large collection of *general* conceptual content than a small collection of more *specific* conceptual content.

The number of documents may not be the main issue for our purpose of using LSA because the corpora used in learning networks are specific to particular sub-domains. It is obvious that in these corpora the dimensionality in the data is less than in broad corpora such as those build on the basis of a complete encyclopaedia. The number of documents needed for LSA is not dependent on the size of the domain (or text database) but on the dimensionality of the domain.

Like corpus size, the “ideal” number of dimensions also is ambiguous. Finding the correct number of dimensions is critical because if it is too small the structure of the data is not captured. If it is too large the latent structure cannot emerge and all unimportant details and sampling error remain. In general, the “magic” number of dimensions was reported between 100 (e.g. Deerwester, Dumais, Furnas, Landauer & Harshman, 1990; Wolfe et al., 1998; Wiemer-Hastings et al., 1999) and 300 (e.g. (Landauer, 2002a; Franceschetti et al., 2001; Olde, Franceschetti, Karnavat, Graeser A.C. & Tutoring Research Group, 2002).

When working with small corpora (< 300 documents), the number of dimensions will obviously be smaller. Recent research showed that it is very well possible to make meaningful use of LSA in small corpora of which the “ideal” dimensionality can be about 40 (Nakov, Valchanova & Angelova, 2003). Below we will discuss some of our own research that yielded similar results. After gathering a collection of documents that grasp the dimensionality of the domain in the best possible way, there are a number of things to do to tidy the corpus. This means that elements that are meaningless to LSA like html code, diacritic tokens and images are removed. Furthermore, it may be desirable to remove redundant text, identical text and spelling errors. Other more specific pre-processing techniques like stopping and stemming are discussed in a separate paragraph.

4.1.3 Document size

Documents for LSA may be as small as individual sentences or as large as essays, articles or web pages. Research reports mention a variety of documents, such as student answers on a test question (e.g. Wild, Stahl, Stermsek & Neumann, 2005), encyclopaedia articles (e.g. Wolfe et al., 1998), parts of textbooks on a certain topic (e.g. Wiemer-Hastings et al., 1999) or student essays (e.g., Rehder et al., 1998). Reports on document size or grain size vary from an average document size of 50 words (e.g. Deerwester et al., 1990) to full articles on a tutoring topic (e.g. Olde et al., 2002). Often full text articles are cut into smaller parts, which are about one paragraph in length. Wiemer-Hastings et al. (1999) state that the paragraph is said to be, in general, a good level of granularity for LSA analysis because a paragraph tends to hold a well-developed coherent idea (Peter Foltz, personal communication, October 1997). This is supported by findings from (Rehder et al., 1998) who found a minimum essay length of 60 words suitable for the purpose of knowledge assessment.

In our own experiment each document was manually split and most of the time matched one paragraph containing information on a single species of monkey or ape.

4.1.4 Stemming

Filtering the data of a corpus can be done by stemming and by stopping. Stemming refers to the parsing of tokens to their semantic stem. Thus, tokens such as "hypothesis", "hypotheses" and "hypothesized" would all be stemmed to a semantic root "hypothes". Stemming has the potential of raising the semantic relevance of the results. For example, the “mountain gorilla” is called the “berglandgorilla” in Dutch, which is parsed as a single, completely different token than “gorilla” In addition to stemming the corpus, stemming a query can also raise semantic relevance. With respect to LSA, stemming alone is reported to show only one to five percent improvement (Dumais, 1992) or even to reduce average correlations with human raters (Wild et al., 2005).

4.1.5 Stopping

Stopping refers to the filtering of noise words, such as "the" and "not" that occur frequently and indiscriminately in the corpus. Noise terms appear throughout any “raw” corpus and do not contribute to the discrimination of documents. From a measurement point of view these terms only add error variance to the corpus. In large corpora that reflect broad domains, such as those based on an encyclopedia or a collection of web sites, the terms to be stopped may be determined by consulting a general list of word frequencies in the written language Wild et al. (2005) indicate that stopping is absolutely necessary. Our results support this conclusion.

5 Methodological considerations

Haley et al. (2005) urge researchers to describe their analyses and data in more detail so that research results can be better compared and understanding of LSA and its further development and refinement are fostered. Elements that need to be described include a description of how the number of singular values retained in the LSA was determined, what frequency measures were used and whether normalization was applied to the document and query vectors. In this section we will concentrate on the most critical issue, which is the determination of the number of singular values.

5.1 Determining the number of singular values

The most important decision in any LSA is the selection of the number of singular values (i.e. the number of dimensions) that will be used to reproduce the data. Dimension reduction itself is not a goal of LSA and it has become accepted practice that for corpora with the size of 5000 documents and above, at least 300 dimensions are used. Larger numbers are not uncommon. As Deerwester et al. (1990) already noted, it is core to LSA to not retain all singular values, because the “latent” factors only emerge in a model with lower dimensionality than the original. The correct choice of the number of singular values is critical in small corpora, where even the number of documents may be less than 300. We clearly need other heuristics than the rule of thumb of 300 dimensions, to determine the number of singular values. As we will demonstrate below, selecting too few factors or too many may have a deteriorating effect on the performance of the LSA.

One method that might be used to decide on the number of singular values is the Scree test (Cattell, 1966) as was used in our example of LSA presented earlier. In this test one visually inspects the data to find the place where a sudden drop occurs in the size of the singular values. This point is used as a cut-off point, beyond which additional singular values are believed to add little more than error to the data. Although the Scree test is generally applicable,

it has the obvious drawback of relying on visual inspection. A second approach is to normalize the document vectors to unitary length, and retain only the singular values that have a length greater than one. This corresponds to common practice in factor analysis of retaining only the eigenvalues larger than one. However, to routinely apply normalization to the data, would cause the risk to loose all information related to the length of documents. The third approach, as proposed by Wiemer-Hastings & Graesser (2000) is to empirically determine the number of dimensions, for example, by selecting the number of singular values that optimises a performance criterion, such as the correlation with ratings by human raters (Wiemer-Hastings et al., 1999).

Van Bruggen et al. (submitted) suggest to combine a performance criterion with a constraint on the amount of variance explained by the singular values. Their reasoning is based on the close connection between singular values and the variance they account for in the term-document matrix. A term-document matrix is, in general, sparse, that is, the matrix contains lots of empty or zero-filled cells. The mean of the cell frequencies will therefore be close to zero and the variance in the matrix will be small as well. For sparse matrices, singular values are closely related to the variance in the matrix. Consider the formula for variance:

$$\frac{\sum x^2}{n} - M^2$$

where x are the cell frequencies, M is the mean frequency in the matrix and n is equal to the number of observations (that is the number of cells in the matrix). In sparse matrices the mean is close to zero and n is a (large) constant, and thus the variance is mainly determined by the sum of squares of the cell frequencies. Since this sum of squares is equal to the sum of squared singular values, the proportion of variance accounted for can be approximated by the sum of squared singular values. Thus, one may select a minimum and a maximum number of singular values that correspond with a bandwidth of variance accounted for. Figure summarizes one of their results. The x-axis represents the number of singular values used; the y-axis is used for correlations as well as explained variance. The ascending curve represents the variance accounted for. The two others represent two performance criteria. The best performance is where the difference between the two is maximum. As predicted, the correlations decrease when more singular values are being used. Once a bandwidth for variance explained is chosen, e.g. 80 to 90%, one can select the number of singular values yielding the optimum performance.

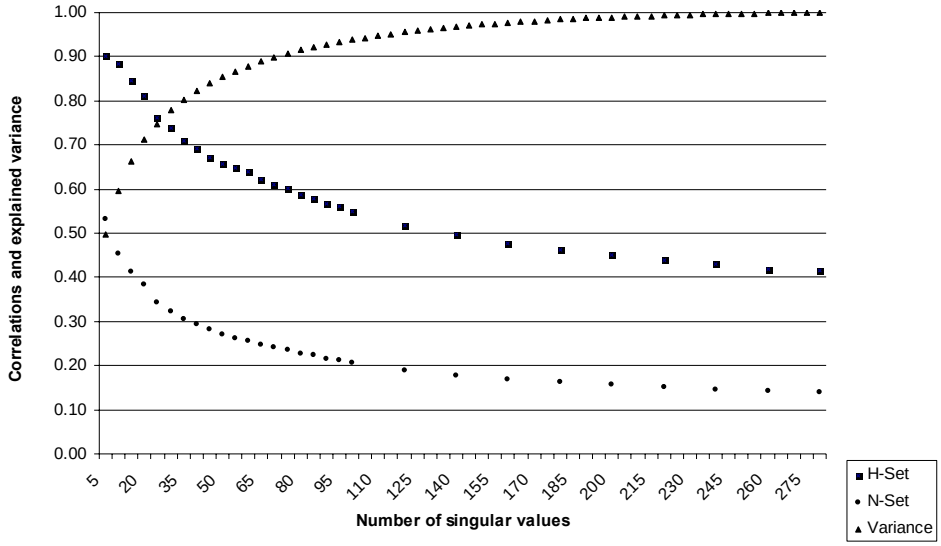

Figure 3. Relation between number of singular values, variance explained and performance

5.2 Weighting: influencing vector length

The raw frequencies in the term–document matrix can be transformed using local and global weighting. An example of a local weighting transformation is a binary transformation in which the frequencies are replaced by a binary value indicating whether a term is present or not. Another local weighting function replaces raw frequencies with their (natural) logarithm. Local weighting can be combined with global weighting that expresses the term occurrence in the corpus. Global weighting is applied to all cells corresponding to a term. Examples here are normalizing transformations, for example to normalize term vectors, and transformations such as inverse frequencies or entropy-measures. According to Berry, Dumais and O'Brien (1994) a combination of a local log transformation with a global entropy weight was found to yield the best performance. These results were not replicated in the extensive parameter testing of Wild (2005) who found that applying inverse document frequencies as global weight increased performance and that none of the local weighting schemes by itself led to better improvement. As the authors indicate, the results are likely to be specific for the corpora studied.

5.3 Measure of similarity

In general, the similarity between documents is computed using the cosine of the angle between the document vectors. There is very little research evidence on the comparison of different similarity measures. Wild et al (2005) reported the best results when they used Spearman's Rho – a similarity measure that uses ordinal data rather than the interval-based data that the others assume.

5.4 Queries

A user query is represented as a vector in the multidimensional space defined by the LSA. The vector is defined as the sum of the terms of the query, weighted by the local and global weights applied to the frequencies in the term-document matrix. Finally the vector is scaled by the

singular values. This query vector is then compared to the document vectors by calculating a similarity measure such as the cosine of the angle between vectors.

5.5 Updating and folding

Adding terms or documents to an existing LSA solution can be done in different ways, all discussed in (Berry et al., 1994). The simplest solution is that of 'folding in' term or document vectors. The procedure for terms is the same as applied for calculating query vectors and document vectors are treated analogous. The new weighted and scaled vectors are added to the existing model. Folding in has the disadvantages that the orthogonality of the model is lost and that new information cannot influence the model. As an alternative to a complete recalculation there are routines to update the solution. We refer the reader to Berry et al. (1994) for detailed descriptions.

6 Evaluation

In order to determine why LSA should or should not be used, an evaluation is provided which reflects the strengths and weaknesses of the technique.

6.1 Strengths of LSA

The main advantage of LSA is that it allows for fairly intelligent operations to be performed by putting in a minimum amount of effort. This is illustrated by its ability to use word co-occurrence data to move words and documents into a reduced dimensionality space where they can be more meaningfully compared to each other. This is fully automatic and does not require the use of metadating, preliminary construed dictionaries, semantic networks, knowledge bases, grammatical syntactic analysers etc.. Automation of processes can alleviate human workload considerably which is an additional advantage.

6.1.1 Strengths through mathematical representation

With respect to other mathematical techniques it has been said by Miller (2003) that LSA concerns inter-word relationships at a deeper level than co-occurrence measures ever could.

Besides saving time, it has been found that the quality of its output can be very high. Essay grading systems which use LSA, consistently outperform those without. For example, over diverse topics, the Intelligent Essay Assessor scores agreed with human experts as accurately as expert scores agreed with each other (Foltz et al., 1999). LSA predicts scores as well as human graders (Landauer & Dumais, 1997); (Wild et al., 2005) and LSA can measure prior knowledge well enough to select appropriate text (Wolfe et al., 1998).

Because LSA uses a mathematical representation of the relations between words in a text and the semantic distance between texts it offers a rapid analysis of large numbers of documents.

As described in section 3.3, LSA has several applications in education. Because of its possibilities to use with respect to dynamic corpora (also see chapter 5) LSA seems ideal for use in learning networks.

6.1.2 Strengths through representation and reduction of complexity of concepts

If the internal representation of semantic similarity is not reduced but constructed in as many dimensions as there are contexts, there would be little practical use for the output of LSA. The strength of LSA is that it represents a corpus in a k dimensional space and thereby reducing the complexity, which makes it possible to improve the estimates of pair wise similarities. It is hereby possible to accurately estimate the similarities among pairs never observed together, by fitting them as best we could into a space of the same dimensionality. For example, research done by Deerwester et al. (1990) shows an LSA model of relations between 60,000 words (30,000 text passages) made with LSA to score on a synonym test for admission in U.S. College to perform as well as the average student who did the test.

6.2 Weaknesses of LSA

The weakness of LSA lies in the empirical determination of computational factors, the computational time that is needed to analyse big corpora, the directionality of LSA and the application in contexts with the emphasis on logic and reasoning.

6.2.1 Empirical determination of computational factors (e.g. singular values)

The number of singular values that offers the best result is no fixed “magic number”. It is very important to determine the right number of dimensions for the amount of success of LSA (Landauer, 2002b).

Often operational criteria are used as a way to determine the ideal number of dimensions, that is, the number of dimensions which delivers the best result is determined (it is probably highly dependent on what kind of result is aimed to be acquired). As seen in chapter 4.1.2, the generally accepted “ideal” number of SV’s lies between 100 and 300 which provides no option if we work with small corpora. We suggest an additional rule of thumb for determining suitable singular values within small-specific corpora: the explained amount of variance.

6.2.2 Computational time for dynamic corpora

As mentioned in section 4.1.1, (Landauer, Laham & Foltz, 1998) state that current computational power is not enough to perform SVD on 100.000’s x 10.000.000’s matrices. (Quesada, Kintsch & Gomez, 2001) also mention the demand for powerful computers to perform necessary analyses. Large matrices like Landauer et al. and Quesada et al. use may not always be needed but computation time may be a problem because of the dynamic nature of material that is often used. This is the case, for example, with transient communities which may require a constant change of corpora. The gravity of this problem is determined by the frequency with which the matrix is updated: if this is once a week or even once in a few days the problem is decreased significantly.

’Directionality’ of knowledge
All LSA input, like an essay for example, is represented by a vector. The direction a vector has, is interpreted as the representation of the quality of the semantic content of that particular piece of input. The cosine does not provide any information about the “directionality” of knowledge because “*it measures relatedness as an unsigned angle in a high-dimensional space*” (Rehder et al., 1998, p.14). This means that:

> […] the essays of two individuals may have the same cosine with an instructional text, but the essay of the first individual may be dissimilar to the text because the individual knows very little about the topic (relative to the text), whereas the essay of the second individual may be dissimilar to the text because the individual knows very much about the topic (relative to the text). (Rehder et al., 1998, p.14).

This can be solved by using a combination of LSA and multi-dimensional scaling (MDS) (Rehder et al. 1998; Carol & Arabie, 1998). MDS is used to (re)calculate subspaces and can compare the distance between different input texts for example to make a distinction between novices and experts. Three methods are available of which one is recommended as most accurate. This method calculates cosines between all pairs of text are to create a detailed Euclidean space. Then, a MDS procedure takes place using a standard procedure (Carol & Arabie, 1998) to create a 10-dimensional space that represents all non-random differences between the cosinusses. This method is effective, but also contains a step of empirical "matching" of the parameters and therefore it is more sensitive to chance and variability.

6.2.3 Reasoning and logics

All LSA models are based on co-occurrence of concepts in documents. The order in which these concepts occur/co-occur is completely ignored. This means that LSA is inadequate to detect logical fallacies. As Wolfe & Goldman (2003) point out, LSA fails to represent domains in which the context determines how sentences should be interpreted. This applies to domains that use metaphorical language, causal reasoning and logically ordered sequences of steps.

Fooling an LSA based essay grader by submitting an essay with just keywords only is possible but will not be a problem. Of course, results will be contaminated, but if a student is capable of writing such an essay, s/he has a very good understanding of the domain and will have proven so (Lemaire & Dessus, 2001).

7 Other applications of LSA

As is shown above, the use of LSA within education is versatile. LSA might also be used to support activities in education other than document retrieval, essay grading etc.. The following paragraphs provide an indication of what these might be. The information provided here is not meant to be exhaustive; many other applications of LSA may be possible.

7.1.1 Community formation, community support and collaboration

LSA can play a role in community forming as well in the support of communities while they are collaborating. Interest as well as expertise areas of members can be based on comparison of document sets that members of the learning communities collect and produce, thus providing shared interests and probably shared goals/aims between community members from the start of the group formation. This 'social matching' process could be supported by the provision of a visualization (as a part of an identity representation) based on the topics people are interested in (comparable with FlickR's visualization of its folksonomy, http://www.flickr.com/photos/tags/). This could help to form informal (sub)groups within the large community of learners, tutors, alumni etc.

Also, while people are collaborating within a community, certain questions will rise. To find the people who are experts on these topics and will probably be able to answer the question, LSA could be useful in the matching of questions and interest areas. Small temporarily sub-groups within the community could be formed, like Kester et al. (2005) do with their transient

communities, to solve the ‘problem’ (=question) based on background and expertise of members.

In both examples of LSA’s usability for communities it becomes easier to find information, like who is doing what and what topics are related. It also helps to match members interests and aims, which can increase motivation to participate. It mainly focuses on matching people to people.

LSA could also help to create a feeling of trust within a group of collaborating people with a specific task and goal within a specific time span, but who don’t know each other and don’t have the opportunity to see each other (a lot). LSA could help to make a certain representation of e.g. interest areas of these people, thus helping other project members to form a mental image of the other person, without spending a lot of time on this image building. This first image is important for the forming of trust, which ultimately has an influence on development of group conflicts and on group performance and interactivity as a whole.

7.1.2 Human Resource Management and task allocation

Other matches could be made with the use of LSA: to position people on the ‘right’ job by matching descriptions of people to a job/role profile, to provide people with the ‘right’ instruction or to provide people with the ‘right’ mentor (‘right’ meaning as personalised to the need/question as possible). Based on a question or the specified need of a student, suggested instructional material could be filtered or a mentor selected based on his/her profile with expertise and interests. In this relatively new field of research and practice, thus far interesting and promising results were obtained within the domain of the army (e.g. Laham et al., 2000), where functional matches between jobs, training and people were made.

7.1.3 Localizing resources

The descriptions above are quite specific applications of LSA. But in both cases, a specific resource is located for a specific user and aim/need/question. In general, LSA is very useful to localize resources: it can compare and determine similarity between two text-based sources, thus determining compatibility. In this way it could be useful in many ways, e.g. localizing experts (within and out of a certain group), localizing documents (e.g. of previous project groups/comparable projects), localizing group of interest and localizing potential sponsors.

7.1.4 Support of assessment and feedback

When project deliverables are largely comparable to previous projects, LSA could also play a role in the final assessment process of projects (like the assessment of airplane landing technique in Quesada (2003)). It could support assessors in their judgement, e.g. to compare the deliverable to previous high quality project deliverables, which were qua problem, domain and content more or less the same. In this way, it could provide a type of framework for the judgement of deliverables.

For students, it could help them to consider alternative perspectives on a topic, by providing feedback and suggestions on related topics while they are working. E.g. suggestions like "previous project groups also considered/mentioned 'x' and 'y' while they were working on this topic".

8 References

Almeida, R. B., & Almeida, V. A. F. (2004). A community-aware search engine. *WWW2004, New York, May 17-22,* 413-421. New York, USA: ACM.

Berry, M. W., Drmac, Z., & Jessup, E. R. (1999). Matrices, vector spaces and information retrieval. *SIAM Review, 41*(2), 335-362.

Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1994, December). Using Linear Algebra for Intelligent Information Retrieval. *SIAM Review, 37*(4), 573-595. Retrieved from http://www.cs.utk.edu/~library/TechReports/1994/ut-cs-94-270.ps.Z

Carrol, J.D. & Arabie, P. (in press). Multidimensional scaling. In M. H. Birnbaum (Ed). *Handbook of Perception and Cognition, Volume 3: Measurement, Judgment and Decision Making*. (pp.179-250). San Diego: Academic Press.

Caron, J. (2000). *Applying LSA to Online Customer Support: A Trial Study.* Retrieved March 13, 2006 from http://www.unidata.ucar.edu/staff/caron/faqo/faqoPaper1.pdf

Cattell, R.B. (1966). The Scree Test for the Number of Factors. *Multivariate Behavioral Research*, 1, 245-276.

Chung-Min, C., Stoffel, N., Post, M., Bassu, D., & Behrens, C. (2001). Telcordia LSI Engine: Implementation and Scalability Issues. *Proceedings of the 11th International Workshop on Research Issues in Data Engineering (RIDE '01), Heidelberg, Germany, April 1-2*.

De Croock, Marcel; Pannekeet, Kees; De Vries, Fred; Sloep, Peter; Van Rosmalen, Peter. (2003) ASA: Agents for Support Activities (project plan). Heerlen: OUNL.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T., & Harshman, R. (1990b). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*, 391-407.

Dessus, P., Lemaire, B., & Vernier, A. (2000). Free-Text Assessment in a Virtual Campus. *Proceedings of the CAPS'2000 conference, Paris, December 13-14*.

Dumais, S. T. (1992). *Enhancing Performance in Latent Semantic Indexing (LSI) Retrieval.* Bellcore.

Dumais, S. T. (1997, April 4). *Using Latent Semantic Indexing (LSI) for information retrieval, information filtering and other things.* Retrieved May 12, 2005, from http://lsa.colorado.edu/papers.html

Foltz, P. W. (1996). Latent semantic analysis for text-based research 25. *Behavior Research Methods, Instruments & Computers, 28*(2), 197-202.

Foltz, P. W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in online writing evaluation with LSA. *Interactive Learning Environments, 8*(2), 111-129. Retrieved from http://www-leibniz.imag.fr/perso/s1/blemaire/public_html/lsa.html

Foltz, P. W., Kintsch, W., & Landauer, T. (1998). The measurement of textual cohesion with latent semantic analysis, *Discourse Processes, 25*(2&3), 285-307.

Foltz, P. W., Laham, D., & Landauer, T. (1999). Automated Essay Scoring: Applications to Educational Technology. *Proceedings of EdMedia '99, Seattle*.

Franceschetti, D. R., Karnavat, A., Marineau, J. M. G. L., Olde, B. A., Terry, B. L., & Graesser, A. C. (2001). Development of physics test corpora for latent semantic analysis. *23th Annual Meeting of the Cognitive Science Society*297-300.

Freeman, J. T., Thompson, B. T., & Cohen, M. S. (2000). Modeling and Diagnosing Domain Knowledge Using Latent Semantic Indexing. *Interactive Learning Environments, 8*(3), 187-209.

Giles, J. T., Wo, L., & Berry, M. W. (2001). GTP (General Text Parser) Software for Text Mining. In *Statistical Data Mining and Knowledge Discovery* (chap. 27).CRC Press. Retrieved from http://www.cs.utk.edu/~berry/papers02/GTPchap.pdf

Graesser, A. C., Hu, X., Olde, B. A., Ventura, M., Olney, A., Louwerse, M., et al. (2005). Implementing Latent Semantic Analysis in Learning Environments with Conversational

Agents and Tutorial Dialog. In W. D. &. S. Gray (Ed.), *24th Annual Meeting of the Cognitive Science Society*37. Mahwah, NJ: Erlbaum.

Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. *Uncertainty in Artificial Intelligence*. Retrieved March 9, 2006,

Kester, L., Sloep, P., Brouns, F., Van Rosmalen, P. De Vries, F. De Croock, M. Koper, R. (submitted). Enhancing Social Interaction and Spreading Tutor Responsibilities in Bottom-Up Organized Learning Networks.

Kanejiya, D., Kumar, A., & Prasad, S. (2003). Automatic Evaluation of Students' Answers using Syntactically Enhanced LSA. *Human Language Technology Conference (HLT-NAACL 2003) Workshop on Building Educational Applications using NLP.* Edmonton, Canada.

Kintsch, E., Steinhart, D., Stahl, G., LSA research group, Matthews, C., & Lamb, R. (2000). Developing summarization skills through the use of LSA-based feedback 30. *Interactive Learning Environments, 8*(2), 87-109.

Kintsch, W. (2002b). The Potential of Latent Semantic Analysis for Machine Grading of Clinical Case Summaries. *Journal of Biomedical Informatics, 35*, 3-7.

Kintsch, W. (1998). The representation of knowledge in minds and machines. *International Journal of Psychology, 33*(6), 411-420.

Kintsch, W. (2002a). On the notions of theme and topic in psychological process models of text comprehension. In M. Louwerse & W. van Peer (Eds.), *Thematics: Interdisciplinary studies* (pp. 157-170). Amsterdam, Netherlands: John Benjamins Publishing Company.

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

Kintsch, W., & Bowles, A. (2002). Metaphor comprehension: What makes a metaphor difficult to understand? *Metaphor and Symbol, 4*(17), 249-262. Retrieved from http://lsa.colorado.edu/papers.html

Koper, R., van Bruggen, J., Rusman, E., & Giesbers, B. (2005). *Learning Technology Development Programme - Positioning in learning networks*.

Laham, D. (1997). Latent Semantic Analysis approaches to categorization. *Proceedings of the 19th annual meeting of the Cognitive Science Society,* 979.

Laham, D., Bennett, W., & Derr, M. (2002). *Latent Semantic Analysis for Career Field Analysis and Information Operations.* Retrieved January 24, 2006, from http://www.k-a-t.com/papers/ab-careerField2002.shtml

Laham, D., Bennett, W., & Landauer, T. K. (2000). An LSA-Based Software Tool for Matching Jobs, People, and Instruction. *Interactive Learning Environments, 8*, 171-185.

Landauer, T. K. (9-8-2002a). Applications of Latent Semantic Analysis. *24th Annual Meeting of the Cognitive Science Society*.

Landauer, T. K. (2002b). On the computational basis of learning and cognition: Arguments from LSA. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory, Vol. 41* (rep. No. 2002-10361-005, pp. 43-84). San Diego, CA, US: Academic Press.

Landauer, T. K., Laham, D., & Foltz, P. W. Jordan, M. I., Kearns, M. J., & Solla, S. A. (Eds.). (1998). Learning human-like knowledge by Singular Value Decomposition: A progress report. *Advances in Neural Information Processing Systems, 10*, 45-51. Retrieved from http://lsa.colorado.edu/papers.html

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2-3), 259-284.

Lemaire, B., & Dessus, P. (2001). A System to Assess the Semantic Content of Student Essays. *Journal of Educational Computing Research, 24*(3), 305-320.

Letsche, T. A. (1997). Large-Scale Information Retrieval with Latent Semantic Indexing. *Informatics and Computer Science, 100*, 105-137. Retrieved from http://www.cs.utk.edu/~berry/lsi++/

Magliano, J. P., Wiemer-Hastings, K., Millis, K. K., Munoz, B. D., & McNamara, D. (2002). Using latent semantic analysis to assess reader strategies. *Behavior Research Methods, Instruments & Computers, 34*(2), 181-188.

Miller, T. (2003). Essay assessment with latent semantic analysis. *Journal of Educational Computing Research, 29*(4), 495-512.

Nakov, P., Valchanova, E., & Angelova, G. (2003). Towards deeper understanding of the LSA Performance. *Recent Advances in Natural Language processing (RANLP'2003)*311-318.

Olde, B. A., Franceschetti, D. R., Karnavat, A., Graeser A.C., & Tutoring Research Group. (2002). The Right Stuff: Do You Need to Sanitize Your Corpus When Using Latent Semantic Analysis. *24th Annual Meeting of the Cognitive Science Society*708-713. Fairfax.

Quesada, J. F. (2003). Introduction to Latent Semantic Analysis and Latent Problem Solving Analysis. In *Latent Problem Solving Analysis (LPSA): A computational theory of representation in complex, dynamic problem solving tasks* (pp. 22-35) (chap. 2). Retrieved from http://www.andrew.cmu.edu/user/jquesada//dissertation/

Quesada, J. F., Kintsch, W., & Gomez, E. (2001). A Computational Theory of Complex Problem Solving Using the Vector Space Model (part I): Latent Semantic Analysis, Through the Path of Thousands of Ants. *Cognitive research with Microworlds, 43*(84), 117-131. Retrieved from http://lsa.colorado.edu/papers.html

Rehder, B., Schreiner, M. E., Wolfe, M. B., Laham, D., Landauer, T. K., & Kintsch, W. (1998). Using latent semantic analysis to assess knowledge: some technical considerations *Discourse Processes, 25*, 337-354.

Rosario, B. (2000). *Latent Semantic Indexing: An Overview* (INFOSYS 240). Retrieved October 10, 2005, from http://www.sims.berkeley.edu/~rosario/projects/LSI.pdf

Stahl, G. Allowing learners to be articulate: incorporating automated text evaluation into collaborative software environments*. Proposal to the McDonnell foundation,* 1997.

Turney, P. D., Litmann, M. L., Bigham, J., & Shnayder, V. (2003). Combining Independent Modules to Solve Multiple-Choice Synonym and Analogy Problems. In G. Angelova, K. Bontcheva, R. Mitkov & Nicolov, N. (Eds.), *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-03), Borovets, Bulgaria, september 10-12, 2003,* 482-389.

van Bruggen, J., Sloep, P., van Rosmalen, P., Brouns, F., Vogten, H., Koper, R., & Tattersall, C. (2004). Latent semantic analysis as a tool for learner positioning in learning networks for lifelong learning. *British Journal of Educational Technology, 35*(6), 729-738.

Van Bruggen, J. M., Rusman, E., Giesbers, B. & Koper, R. (submitted). Latent semantic analysis of small-scale corpora for positioning in learning networks.

Wade-Stein, D., & Kintsch, E. (2003). *Summary Street: Interactive Computer Support for Writing* (03-01(2003)). Colorado: University of Colorado. (Sectie W).Retrieved May 12, 2005, from http://www-leibniz.imag.fr/perso/s1/blemaire/public_html/lsa.html

Wiemer-Hastings, P. (1999). How latent is Latent Semantic Analysis? 29. *Proceedings of the Sixteenth International Joint Congress on Artificial Intelligence,* 932-937*.* San Francisco: Morgan Kaufmann.

Wiemer-Hastings, P., & Graesser, A. C. (2000). Select-a-Kibitzer: a computer tool that gives meaningful feedback on student compositions, *Interactive Learning Environments, 8*(2), 149-169.

Wiemer-Hastings, P., Wiemer-Hastings, K., & Graesser, A. C. (1999). Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis. In S. P. Lajoie & Vivet M. (Eds.), *Artificial Intelligence in Education.* Amsterdam: IOS Press. 535-542.

Wiemer-Hastings, P., & Zipitria, I. (2001). Rules for syntax, vectors for semantics. *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society,* 1112-1117*.* Mahwah,NJ: Lawrence Erlbaum Associates.

Wild, F., Stahl, C., Stermsek, G., & Neumann, G. (2005). Parameters driving effectiveness of automated essay scoring with LSA. *Proc. of the 9th International Computer Assisted Assessment Conference (CAA), Loughborough, July, 2005,* 485-494.

Wild, F., Stahl, C., Stermsek, G., Penya, Y., & Neumann, G. (2005). Factors influencing effectiveness in automated essay scoring with LSA. *Proc. of the 12th International*

Conference on Artificial Intelligence in Education (AIED), Amsterdam, july, 2005 IOS Press. 485-494.

Wolfe, M. B. W., & Goldman, S. R. (2003). Use of latent semantic analysis for predicting psychological phenomena: Two issues and proposed solutions. *Behavior Research Methods, Instruments & Computers, 35*(1), 22-31.

Wolfe, M. B. W., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W., et al. (1998). Learning from text: Matching readers and texts by latent semantic analysis. *Discourse Processes, 25*(2-3), 309-336.

Yu, C., Cuadrado, J., Ceglowski, M., & Payne, J. S. (4-2-2005). *Patterns in Unstructured Data - Discovery, Aggregation, and Visualization.* Retrieved May 12, 2005, from http://research.nitle.org/lsi/cover_page.htm

Zampa, V., & Lemaire, B. (2002, October 15). Latent Semantic Analysis for User Modelling. *Journal of Intelligent Information Systems, 18*(1), 15-30. Retrieved from http://www-leibniz.imag.fr/perso/s1/blemaire/public_html/lsa.html

9 General bibliography on LSA

The following bibliography was generated on July 6th, 2006, from our literature database on LSA. The (Reference Manager®) database, and a RIS version (*.txt format, can be imported in other applications like Endnote) are available at the secretary of the Research Technology Development Programme of the Educational Technology Expertise Center.

Alaniz, A., Graca Campos, M., & Antonio, J. (2006). An infrastructure for Open Latent Semantic Linking. In ACM (Ed.), *HT'02* (pp. 107-116).

Almeida, R. B. & Almeida, V. A. F. (2004). A community-aware search engine. In *WWW2004* (pp. 413-421). New York, USA: ACM.

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. New York: ACM Press.

Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1994). Using Linear Algebra for Intelligent Information Retrieval. *SIAM Review, 37,* 573-595.

Berry, M. W., Drmac, Z., & Jessup, E. R. (1999). Matrices, vector spaces and information retrieval. *SIAM Review, 41,* 335-362.

Abstract: This paper shows how fundamental mathematical concepts from linear algebra can be used to manage and index large text collections.

Boone, G. N. (2000). *Extreme Dimensionality Reduction for Text Learning: Cluster-generated Feature Spaces.* Georgia Institute of Technology.

Bouma, G. & Klein, E. H. (2001). Volkskrant database and practicum.

Ref Type: Data File

Burstein, J., Kaplan, R., Wolff, S., & Lu, C. (1996). Using lexical semantic techniques to classify free-responses. In: Proceedings from the SIGLEX19666 workshop, Annual meeting of the Association of Computational Linguistics. University of California, Santa Cruz.

Ref Type: Unpub. contribution symposium

Notes: Beschrijft twee experimenten waarin wordt gekeken in welke mate beoordelaars en LSA overeenstemming in beoordeling van de inhoud van essays en in welke mate ze de scores op een objectieve test kunnen voorspellen. In **experiment 1** produceerden 94 undergraduates

essays van ongeveer 250 woorden over de anatomie, functie en doel van het menselijk hart. Twee professionele lezers van ETS scoorden de essays op een vijfpuntsschaal, nadat ze kennis hadden genomen van achtergrondmateriaal en hadden bepaald welke inhoud de essays dienden te bevatten. Stundenten maakten tevens een "40 point" toets over de stof. LSA getraind op 27 artikelen: 94 dimensies voor 830 zinnen en 3034 unieke woorden. Ieder essay kreeg een vector berekend op basis van het gemiddelde van de woordvectoren. LSA werd op twee manieren gebruikt: methode 1: cos target essays berekend met alle andere essays. Het doelessay kreeg vervolgens het gemiddelde van de beoordelingen die de tien dichtsgelegen essays hadden gekregen van de beoordelaars. Een tweede maat was de lengte van het essayvector. Resultaten: correlatie beoordelaars: .77; LSA met beoordelaars .77, .68 ; gemiddelde: .77. Beoordelaars en objectieve toets: r=.70, LSA: r=.81. Methode 2: beoordeling dit keer door het essay te vergelijken met een expert tekst uit een studieboek. Resultaten: LSA en beoordelaars: r=.64, .71, gemiddeld: .72. LSA en extern criterium: 77. Verdere analyse liet zien dat zowel technische als niet-technische begrippen bijdroegen aan de cos (en scores), maar dat alleen de vectorlengte van de technische woorden bijdroeg aan de voorspelling van het criterium. In **Experiment 2** vervaardigden 273 studenten die een inleiding psychologie volgden een essay (tien munuten beschikbaar) over een van drie topics (afasie, operant conditioneren, binding bij kinderen). Beoordeling door twee deskundigen. LSA getraind op het gebruikte boek: 4904 alinea's met 19153 unieke woorden, geen stoplist gebruikt: 1500 dimensies bleek hoogste correlaties op te leveren.Nogal wat verschillen tussen de gebruikte teksten (beoordelaars kwamen weinig overeen bij tekst binding). Gemiddelde r over alle teksten: beoordelaars onderling: .65; LSA en beoordelaars gemiddeld: .64.

Cardoso-Cachopo, A. & Oliveira, A. L. (2003). An Empirical Comparison of Text Categorization Methods. In M. A. Nascimento, E. S. de Moura, & A. L. Oliveira (Eds.), *String Processing and Information Retrieval: 10th International Symposium* (pp. 183-196).

Carrol, J.D. & Arabie, P. (in press). Multidimensional scaling. In M. H. Birnbaum (Ed). Handbook of Perception and Cognition, Volume 3: Measurement, Judgment and Decision Making. (pp.179-250). San Diego: Academic Press.

Caron, J. (2000). Applying LSA to Online Customer Support: A Trial Study. 13-3-2006. Ref Type: Unpublished Work

Abstract: In this work, I report on a prototype system for technical support called the Frequently Asked Question Organizer (FAQO). This application enables technical support personnel to construct a knowledge base from email archives and other existing documents. Users can query the knowledge base using natural-language questions in order to find relevant documents. The prototype uses Latent Semantic Analysis (LSA) for query matching.
A technical-support person at the Unidata Program Center tested the application for three weeks by querying the database with all technical questions that came in to him during that period, and rating the returned documents. About half the time emails were found that could help answer the question, and FAQO was found to be superior to the keyword-search tool previously used. Other experiments in matching questions with answers are reported here, along with preliminary precision/recall results that varied some of the key parameters of the LSA algorithm. The main contribution is the engineering of the application itself, which is designed for ease-ofuse and for future modification and enhancements.

Cattell, R. B. (1966). The Scree Test for the Number of Factors. *Multivariate Behavioral Research, 1,* 245-276.

Cherniavsky, J. C. & Soloway, E. (2002). A Survey of Research Questions for Intelligent Information Systems in Education (Editorial). *Journal of Intelligent Information Systems, 18,* 5-14.

Chua, T.-S. (2002). Complementary Content. IEEE Multimedia [On-line].

Chung-Min, C., Stoffel, N., Post, M., Bassu, D., & Behrens, C. (2001). Telcordia LSI Engine: Implementation and Scalability Issues. In *Proceedings of the 11th International Workshop on Research Issues in Data Engineering (RIDE '01)*.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis
26. *Journal of the American Society for Information Science, 41,* 391-407.

Dessus, P., Lemaire, B., & Vernier, A. (2000). Free-Text Assessment in a Virtual Campus. In *Proceedings of the CAPS'2000 conference*.

Dong, A. (2004). Quantifying Coherent Thinking in Design: A Computational Linguistics Approach. In J. Gero (Ed.), *Design Computing and Cognition '04* Dordrecht, The Netherlands: Kluwer Academic

Publishers.

Abstract: Design team conversations reveal their thinking patterns andbehaviour because participants must communicate their thoughts toothers through verbalcommunication. This article describes a methodbased on latent semantic analysis for measuring the coherence of theircommunication in aconversation mode and how this measurement alsoreveals patterns of interrelations between an individual's ideas and thegroup's ideas. While similarstudies have been done on designdocumentation, it was unclear whether computational techniques thathave been applied to communication in textcould be successfullyapplied to communication in a conversational mode. Transcripts offour engineering/product design teams communicating in asynchronous, conversational mode during a design session werestudied. Based on the empirical results and the proposition that ateam's verbal communication offers afairly direct path to theirthinking processes, the article proposes the link between coherentconversations and coherent thinking.

Dumais, S. T. (1992). *Enhancing Performance in Latent Semantic Indexing (LSI) Retrieval.* Bellcore.

Dumais, S. T., Landauer, T. K., & Littman, M. L. (22-8-1996). Automatic cross language information retrieval using Latent Semantic Analysis. SigIR Multilingual IR Workshop.

Ref Type: Unpub. contribution symposium

Notes: document consists of slides

Dumais, S. T. (1997). Using Latent Semantic Indexing (LSI) for information retrieval, information filtering and other things. [On-line]. Available: http://lsa.colorado.edu/papers.html

Dunn, J. C., Almeida, O. P., Barclay, L., Waterreus, A., & Flicker, L. (2002). Latent semantic analysis: A new method to measure prose recall. *Journal of Clinical & Experimental Neuropsychology, 24,* 26-35.

Abstract: Compared traditional methods of scoring the Logical Memory test of the Wechsler Memory Scale-III with a new method based on Latent Semantic Analysis (LSA). LSA represents texts as vectors in a high-dimensional semantic space and the similarity of any 2 texts is measured by the cosine of the angle between their respective vectors. The Logical Memory test was administered to a sample of 72 elderly individuals (aged 64-92 yrs), 14 of whom (aged 72-86 yrs) were classified as cognitively impaired by the Mini-Mental State Examination. The results

show that LSA was at least as valid and sensitive as traditional measures. Partial correlations between prose recall measures and measures of cognitive function indicated that LSA explained all the relationship between Logical Memory and general cognitive function. This suggests that LSA may serve as an improved measure of prose recall. (PsycINFO Database Record (c) 2002 APA, all rights reserved)

Fletcher, C. R. & Linzie, B. (1998). Motive and opportunity: Some comments on LSA, HAL, KDC, and principal components. *Discourse Processes, 25,* 355-361.

Abstract: Comments on the articles in this special issue, which describe new, highly quantitative techniques for exploring readers' mental representations of discourse. They make a strong case that techniques such as Knowledge Diagraph Contribution analysis, Latent Semantic Analysis, the Hyperspace Analog to Language, harmony maximization, and principal components analysis can be used to solve interesting theoretical and applied problems. As a result, researchers in the area of discourse comprehension and the mental representation of discourse should feel motivated to adopt these techniques. In some cases their ability to do so will be enhanced by the availability of well-documented, easy-to-use computer software, complete with demonstrations and examples. In other cases, they are likely to be stymied by the unavailability of software and support. (PsycINFO Database Record (c) 2002 APA, all rights reserved)

Foltz, P. W. (1996). Latent semantic analysis for text-based research

25. *Behavior Research Methods, Instruments & Computers, 28,* 197-202.

Abstract: Describes 3 experiments that illustrate how latent semantic analysis (LSA), an automatic statistical model of word usage that permits comparisons of semantic similarity between pieces of textual information, may be used in text-based research, in 28 college students. Experiments 1 and 2 involved methods for analyzing an S's essay to determine from what text the S learned the information and for grading how much relevant information was cited in the essay. Experiment 3 involved an approach to using LSA to measure the coherence and comprehension of texts. Results show that LSA was a successful approach for predicting the source of an S's knowledge on the basis of what the S wrote and for characterizing the quality of essays. It was also useful in measuring textual coherence. Thus, LSA is an automatic and fast method that permits quick measurements of the semantic similarity between pieces of textual information. (PsycINFO Database Record (c) 2002 APA, all rights reserved)

Notes: Dit artikel bespreekt drie experimenten die illustreren hoe LSA gebruikt kan worden in tekst-gebaseerd onderzoek: twee experimenten om op basis van essays te beoordelen wat iemand heeft geleerd en dat te waarderen en een om tekstuele cohesie te onderzoeken. Foltz rapporteert drie onderzoeken, waarvan de eerste er op was gericht om te achterhalen op welke documenten studenten zich baseren bij het vervaardigen van een samenvatting (heranalyse data Britt, Perfetti, Rouet & Mason). 24 studenten lazen 21 teksten over de interventie door de VS in Panama en schreven een essay over de vraag of de interventie gerechtvaardigd was.Voor de LSA analyse werd het systeem ook gevoed met algemene artikelen over Panama. Twee raters beoordeelden op welke documenten studenten zich baseerden (een gezamenlijk document is overeenstemming). De raters haalden zo (slechts) 63% overeenstemming. De overeenstemming met LSA bedroeg 56% en 49%.

In een tweede experiment lazen vier doctoraal studenten Geschiedenis de 21 bronteksten en beoordeelden de 24 essays op a) welke informatie er in was verwerkt en b) de kwaliteit van het essay. Ze selecteerden bovendien de tien belangrijkste zinnen uit de bronteksten. Er werden twee methoden gebruikt om LSA te laten beoordelen. Bij de eerste methode werd iedere zin in het essay vergeleken met iedere zin in de bronteksten en het gemiddelde werd genomen van de cosinussen met de best overeenkomende bronzinnen. (maat voor plagiaat of rote recall, aldus Foltz). Voor de tweede maat werd de dezelfde procedure gebruikt, maar nu werd vergeleken met de tien zinnen die de experts het meest belangrijk vonden. De correlaties tussen beoordelaars liepen uiteen van .381, .582 tot .768 (een minder ervaren beoordelaar vielo wat buiten de boot). De LSA beoordeling op basis van tekstoverlap correleerde van .317 tot .552 met de revaren beoordelaars. De correlatie van de LSA beoordeling op basis van expert model liep uiteen van .384 tot .626.

Een derde experiment richte zich op tekstuele coherentie.

Foltz, P. W., Kintsch, W., & Landauer, T. (1998). The measurement of textual cohesion with latent semantic analysis. *Discourse Processes, 25,* 285-307.

Abstract: Latent Semantic Analysis is used as a technique for measuring the coherence of texts. By comparing the vectors for two adjoining segments of text in a high-dimensional semantic space, the method provides a characterization of the degree of semantic relatedness between the segments. We illustrate the approach for predicting coherence through re-analyzing sets of

texts from two studies that manipulated the coherence of texts and assessed readers' comprehension. The results indicate that the method is able to predict the effect of text coherence on comprehension and is more effective than simple term-term overlap measures. In this manner, LSA can be applied as an automated method that produces coherence predictions similar to propositional modeling. We describe additional studies investigating the application of LSA to analyzing discourse structure and examine the potential of LSA as a psychological model of coherence effects in text comprehension.

Notes: Gebruikte techniek is vectorcorrelatie van telkens twee opeenvolgende segmenten hier zinnen (nb LSA pakt zowel de semantische cohesie als de expliciete coreferenties). Heranalyse van twee eerder gebruikte teksten: Britton & Gulgoz, waar coherentie werd gemanipuleerd door bepaalde inhoudswoorden te herhalen en McNamara et al waar coherentie werd gemanipuleerd door woorden en frasen te vervangen door synoniemen. Britton & Gulgoz maakten drie aangepast versies: *Principled* waarin coherentie-gaten (obv propositionele analyse) werden gedicht door coreferenties in te voegen; *Heuristic* waarbij de tekst met de hand werd geredigeerd om de best mogelijke versie te maken en *Readability* waarin werd geoptimaliseerd op readability maten.Britton en Gulgoz onderzochten het effect op recall van proposities, props/minuut (=efficiency) en inference (mc test). De correlaties met de LSA maten bleken in de heranalyse zeer hoog (r = .98, .99, 1.00). Eigenlijk niet zo verrassend omdat LSA erg gevoelig is voor overlap van woorden. De heranalyse van de McNamara gegevens is interessanter. In dat onderzoek werd vooral gewerkt met synoniemen en werd zowel de locale als de macrocoherentie gemanipuleerd (2x2 design laag - hoog, locaal vs macro). McNamara et al vonden dat leerlingen met een lage voorkennis vooral profiteerden van de maximaal coherente tekst en dat leerlingen met een hoge voorkennis vooral baat hadden bij de laag-coherente tekst. Heranalyse hier is gebaseerd op leerlingen met lage voorkennis.LSA maten correleerden laag met woordoverlap maten, maar hoog met overall posttest scores (.94), maar vooral met de tekst-gebaseerde vragen (r=.98). Flesch tests lieten geen verschillen tussen de teksten zien.

Foltz, P. W. & Wells, A. D. (1999). Automatically deriving readers' knowledge structures from texts. *Behavior Research Methods, Instruments & Computers, 31,* 208-214.

Abstract: Latent semantic analysis (LSA) serves as both a theory and a method for representing the meaning of words based on a statistical analysis of their contextual usage (P. W. Foltz, 1996;

T. K. Landauer and S. T. Dumais, 1997). In 2 experiments, in the domains of psychology and history, with 21 and 40 Ss, respectively, the authors compared the representation of readers' knowledge structures of information learned from texts with the representation generated by LSA. Results indicated that LSA's representation is similar to readers' representations. In addition, the degree to which the reader's representation is similar to LSA's representation is indicative of the amount of knowledge the reader has acquired and of the reader's reading ability. This approach has implications both as a model of learning from text and as a practical tool for performing knowledge assessment. (PsycINFO Database Record (c) 2002 APA, all rights reserved)

Foltz, P. W., Laham, D., & Landauer, T. (1999). Automated Essay Scoring: Applications to Educational Technology. In *Proceedings of EdMedia '99*.

Abstract: The Intelligent Essay Assessor (IEA) is a set of software tools for scoring the quality of essay content. The IEA uses Latent Semantic Analysis (LSA), which is both a computational model of human knowledge representation and a method for extracting semantic similarity of words and passages from text. Simulations of psycholinguistic phenomena show that LSA reflects similarities of human meaning effectively. To assess essay quality, LSA is first trained on domain-representative text. Then student essays are characterized by LSA representations of the meaning of their contained words and compared with essays of known quality on degree of conceptual relevance and amount of relevant content. Over many diverse topics, the IEA scores agreed with human experts as accurately as expert scores agreed with each other. Implications are discussed for incorporating automatic essay scoring in more general forms of educational technology.

Foltz, P. W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in online writing evaluation with LSA. *Interactive Learning Environments, 8,* 111-129.

Abstract: This paper describes tests of an automated essay grader and critic that uses Latent Semantic Analysis. Several methods which score the quality of the content in essays are described and tested. These methods are compared against human scores for the essays and the results show that LSA can score as accurately as the humans. Finally, we describe the implementation of the essay grader/critic in an undergraduate course. The outcome showed that students could write and revise their essays online, resulting in improved essays. Implications are

discussed for the use of the technology in undergraduate courses and how it can provide an effective approach to incorporating more writing both in and outside of the classroom.

Franceschetti, D. R., Karnavat, A., Marineau, J. M. G. L., Olde, B. A., Terry, B. L., & Graesser, A. C. (2001). Development of physics test corpora for latent semantic analysis. In *23th Annual Meeting of the Cognitive Science Society* (pp. 297-300).

Abstract: Student responses to qualitative physics questions were analyzed with latent semantic analysis (LSA), using different text corpora. Physics potentially has a number of distinctive characteristics that are not encountered in many other knowledge domains. Physics texts exist at a variety of levels and typically involve an integrated presentation of text, figures and equations. We explore the adequacy of several text corpora and report results on vector lengths and correlations between key terms in elementary mechanics. The results suggest that a carefully constructed smaller corpus may provide a more accurate representation of fundamental physical concepts than a much larger one.

Freeman, J. T., Thompson, B. T., & Cohen, M. S. (2000). Modeling and Diagnosing Domain Knowledge Using Latent Semantic Indexing. *Interactive Learning Environments, 8,* 187-209.

Giles, J. T., Wo, L., & Berry, M. W. (2001). GTP (General Text Parser) Software for Text Mining. In *Statistical Data Mining and Knowledge Discovery* (CRC Press.

Graesser, A., Wiemer-Hastings, Katja, Wiemer-Hastings, P., and Kreuz, R. (1999). AutoTutor: A simulation of a human tutor. *Journal of Cognitive Systems Research,* 35-51.

Graesser, A. C., Hu, X., Olde, B. A., Ventura, M., Olney, A., Louwerse, M. et al. (2005). Implementing Latent Semantic Analysis in Learning Environments with Conversational Agents and Tutorial Dialog. In W. D. &. S. Gray (Ed.), *24th Annual Meeting of the Cognitive Science Society* (pp. 37). Mahwah, NJ: Erlbaum.

Haley, D., Thomas, P., Nuseibeh, B., Taylor, J., & Lefrere, P. (2003). E-Assessment using Latent Semantic Analysis. In *roceedings of the 3rd International LeGE-WG Workshop.*

Abstract: E-assessment is an important component of e-learning and e-qualification. Formative and summative assessment serve different purposes and both types of evaluation are critical to the pedagogical process. While students are studying, practicing, working, or revising, formative

assessment provides direction, focus, and guidance. Summative assessment provides the means to evaluate a learner's achievement and communicate that achievement to interested parties. Latent Semantic Analysis (LSA) is a statistical method for inferring meaning from a text. Applications based on LSA exist that provide both summative and formative assessment of a learner's work. However, the huge computational needs are a major problem with this promising technique. This paper explains how LSA works, describes the breadth of existing applications using LSA, explains how LSA is particularly suited to e-assessment, and proposes research to exploit the potential computational power of the Grid to overcome one of LSA's drawbacks

Haley, D. T., Thomas, P., Roeck de, A., & Petre, M. (2005). *A Research Taxonomy for Latent Semantic Analysis-Based Educational Applications* (Rep. No. 2005/09).

Abstract: The paper presents a taxonomy that summarises and highlights the major research into Latent Semantic Analysis (LSA) based educational applications. The taxonomy identifies five main research themes and emphasises the point that even after more than 15 years of research, much is left to be discovered to bring the LSA theory to maturity. The paper provides a framework for LSA researchers to publish their results in a format that is comprehensive, relatively compact, and useful to other researchers.

Notes: Nuttig literatuuroverzicht met achterin een uitgebreide tabel waarin belangrijke bronnen tegen elkaar af worden gezet aan de hand van doel, innovatie, belangrijkste resultaten plus een tabel met de technische details van de bronnen (o.a. compositie, onderwerp, aantal woorden en documenten)

Hearst, M. (2000). The debate on automated essay grading. *IEEE Intelligent Systems, 5,* 22-37.

Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. *Uncertainty in Artificial Intelligence*.

Abstract: Probabilistic Latent Semantic Analysis is a novel statistical technique for the analysis of two{mode and co-occurrence data, which has applications in information retrieval and ltering, natural language processing, machine learning from text, and in related areas. Compared to standard Latent Semantic Analysis which stems from linear algebra and performs a Singular Value Decomposition of co-occurrence tables, the proposed method is based on a mixture decomposition derived from a latent class model. This results in a more principled approach which has a solid foundation in statistics. In order to avoid overtting, we propose a widely applicable generalization of maximum likelihood model tting by tempered EM. Our approach

yields substantial and consistent improvements over Latent Semantic Analysis in a number of experiments.

Hofmann, T. & Fisher, D. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning, 42,* 177-196.

Abstract: Presents probabilistic latent semantic analysis as a means of identifying and distinguishing between different contexts of word usage in document collections and text corpora. In contrast with latent semantic analysis, which stems from linear algebra and performs a singular value decomposition of co-occurrence tables, the proposed technique uses a generative latent class model to perform a probabilistic mixture decomposition. This results in a more principled approach with a solid foundation in statistical inference. The use is proposed of a temperature controlled version of the expectation maximization algorithm for model fitting, which has shown excellent performance in practice. Probabilistic latent semantic analysis has many applications, most prominently in information retrieval, natural language processing, and machine learning from text. Perplexity results are presented for different types of text and linguistic data collections and an application in automated document indexing is discussed. The results indicate substantial and consistent improvements of the probabilistic method over standard latent semantic analysis. (PsycINFO Database Record (c) 2002 APA, all rights reserved)

Hu, X., Cai, Z., Franceschetti, D. R., Penumatsa, P., Graesser, A. C., Louwerse, M. M. et al. (2005). LSA: First dimension and dimensional weighting. 14-3-2006.

Ref Type: Unpublished Work

Husbands, P., Simon, H., & Ding, C. (2000). On the Use of Singular Value Decomposition for Text Retrieval. In M. Berry (Ed.), *Proc. of SIAM Comp. Info. Retrieval Workshop*.

Hüning, M. (2005). TextStat Simple text analysis tool [Computer software]. Berlin: Dutch Linguistics Free University of Berlin.

Kanejiya, D., Kumar, A., & Prasad, S. (2003). Automatic Evaluation of Students' Answers using Syntactically Enhanced LSA. In *Human Language Technology Conference (HLT-NAACL 2003) Workshop on Building Educational Applications using NLP* Edmonton, Canada.

Abstract: Latent semantic analysis (LSA) has been used in several intelligent tutoring systems(ITS's)

for assessing students' learning by evaluat-ing their answers to questions in the tutoring domain. It is based on word-document co-occurrence statistics in the training corpus and a dimensionality reduction technique. How-ever, it doesn't consider the word-order or syntactic information, which can improve the knowledge representation and therefore lead to better performance of an ITS. We present here an approach called Syntactically Enhanced LSA (SELSA) which generalizes LSA by con-sidering a word along with its syntactic neighborhood given by the part-of-speech tag of its preceding word, as a unit of knowledge representation. The experimental results on Auto-Tutor
task to evaluate students' answers to ba-sic computer science questions by SELSA and its comparison with LSA are presented in terms of several cognitive measures. SELSA is able to correctly evaluate a few more answers than LSA but is having less correlation with human evaluators than LSA has. It also provides bet-ter discrimination of syntactic-semantic knowl-edge representation than LSA.

Kintsch, E., Steinhart, D., Stahl, G., LSA research group, Matthews, C., & Lamb, R. (2000). Developing summarization skills through the use of LSA-based feedback
30. *Interactive Learning Environments, 8,* 87-109.

Notes: Bespreekt de ontwikeling van State the Essence en Summary Street, twee systemen die op basis van LSA samenvattingen beoordelen en gerichte feedback geven. Interessante aspecten: geen gouden standaard, maar docenten markeren de topics in de bronteksten die ook in de samenvattingen aanwezig moeten zijn. Er bestaat een 11Mb algemene ruimte voor highschool maar voor technische details moeten afzonderlijke semantische ruimtes worden gemaakt. Voorbeeld 830 documenten met 17688 woorden voor een ruimte waarin de werking van het hart wordt beschreven en 530 documeten met 46951 woorden voor een semantische ruimte over midden-amerikaanse culturen.

Inhoudelijke feedback is gebaseerd op de cos van de samenvattingsvector met de topics T1...Tn. Zodra de cos > empririsch te bepalen drempel kan feedback worden gegeven dat de inhoud niet voldoende is weergegeven. Als de tekst te lang is wordt feedback gegeven die helpt de tekst in te korten. De cos van iedere zin met de samenvatting als geheel wordt berekend. Zodra cos beneden een empirische te bepalen waarde valt, wordt de zin als irrelevant bestempeld. De cos van iedere zin met iedere andere zin wordt berekend. Zodra cos groter is dan een bovenlimiet

wordt de student gevraagd de zinnen te controleren op redundantie en ze weg te gooien of te combineren.

Het artikel bevat veel details over benodigde input; lengte van samenvattingen e.d. Lln lazen 10 teksten over energiebronnen van ieder 2 - 2,5 pagina. Samenvatting van 75 tot 200 woorden. Ander domein: oude beschavingen: drie teksten van 2-2,5 pagina; smenvattingen 200 tot 300 woorden.Bloedsomloop - geen details over onderliggende teksten - samenvattingen van 150-250 woorden gericht op diepe kennis.

In het artikel wordt uitvoerig beschreven hoe men worstelt met het interface en vooral de feedback van het systeem. In de eerste versies zat te veel feedback en te weinig begeleiding. Technisch gezien was een probleem dat de de scores niet betrouwbaar waren en soms op basis van locale variaties opgeblazen werden (vooral zinsgewijze maten bleken hiervoor gevoelig). Herst 1998 werden vergelijkingen met menselijke beoordelaars gemaakt. LSA cosinus samenvatting -bronteksten bleek bij 50 samenvatting een correlatie van 0,64 te heben met de beoordeling door de docent. De correlatie tussen beoordelaars bedroeg 0,69. Tweede test: matching van 119 zinnen op topic. Menselijke beoordelaars vertoonden grote overeenstemming (91,6%). LSA kwam tot een overeenstemming van 84,9% met dene en 83,2% met de tweede beoordelaar. In een derde onderzoek werd gekeken naar effecten op tekstbegrip en kwaliteit van samenvattingen (39 lln, twee samen te vatten teksten over bloedsomloop). Geen verschil gevonden. Interesant detail de beoordeling op een tienpunts-schaal door menselijke beoordelaars correleerden 0,59. Vergelijking van LSA en menselijke beoordelaars op Energy unit liet enkele interessante resultaten zien: gemiddelde correlatie tussen beoordelaars en LSA bedroeg 0.88 voor vier teksten, maar erg laag op de zes andere topics. Niet alleen missing data zijn hiervoor verantwoordelijk: LSA gebruikt een drempelwaarde voor alle topics en sommige daarvan zijn beduidend moeilijker: LSA geeft voor sommige topics hogere beoordelingen dan de docenten.

Verder onderzoek met de nieuwe versie Summary Street liet voor het eerst significante resultaten op leerresultaten zien, voor moeilijke onderwerpen.De onderzoekers concluderen dat de inhoudelijke feedback vooral helpt als de leerlingen geconfronteerd worden met moeilijker taken of moeilijker teksten.

Kintsch, W. (1998). The representation of knowledge in minds and machines. *International Journal of Psychology, 33,* 411-420.

Abstract: Human knowledge can be represented as a propositional network in which the meaning of a node is defined by its position in the network. That is, the relationship between a node and its neighbors determines how this node is used in language understanding and production (i.e., its meaning). The propositions that make up such a network are predicate-argument structures with time and location slots. Schemas, frames, and production rules can be expressed in the same formalism. Implications for this contextual view of meaning are discussed. Since the construction of such a propositional network depends on hand coding and is therefore impractical, an alternative automatic statistical procedure is explored that yields a high-dimensional semantic space. Vectors in this space correspond to nodes in the propositional network, in that the meaning of a vector in the latent semantic analysis space is given by its neighboring vectors in that space. (PsycINFO Database Record (c) 2002 APA, all rights reserved)

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

Abstract: Metaphor comprehension involves an interaction between the meaning of the topic and vehicle terms of the metaphor. Meaning is represented by vectors in a high-dimensional semantic space. Predication modifies the topic vector by merging it with selected features of the vehicle vector. The resulting metaphor vector can be evaluated by comparing it with known landmarks in the semantic space. Thus, metaphorical predication is treated in the present model in exactly the same way as literal predication. Some experimental results concerning metaphor comprehension are simulated within this framework, such as the non-reversibility of metaphors, priming of metaphors with literal statements, and priming of literal statements with metaphors.

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

Abstract: In Latent Semantic Analysis (LSA) the meaning of a word is represented as a vector in a high-dimensional semantic space. Different meanings of a word or different senses of a word are not distinguished. Instead, word senses are appropriately modified as the word is used in

different contexts. In N-VP sentences, the precise meaning of the verb phrase depends on the noun it is combined with. An algorithm is described to adjust the meaning of a predicate as it is applied to different arguments. In forming a sentence meaning, not all features of a predicate are combined with the features of the argument, but only those that are appropriate to the argument. Hence, a different "sense" of a predicate emerges every time it is used in a different context. This predication algorithm is explored in the context of four different semantic problems: metaphor interpretation, causal inferences, similarity judgments, and homonym disambiguation.

Kintsch, W. (2002). On the notions of theme and topic in psychological process models of text comprehension. In M.Louwerse & W. van Peer (Eds.), *Thematics: Interdisciplinary studies* (pp. 157-170). Amsterdam, Netherlands: John Benjamins Publishing Company.

Abstract: Describe the features of theme and topics as a first step in exploring the analogies and correspondences between psychological process models of language understanding and the linguistic notions of theme and topic. The author focuses on a specific process model of text comprehension, the construction-integration theory, which attempts to simulate the computations involved in the construction of a mental representation of a text in human comprehension. The author also discusses the use of latent semantic analysis to simulate human verbal knowledge and the use of this analysis to generate the macrostructure of a text. The author argues that latent semantic analysis allows one to derive a precise mathematical representation of the topic or theme of a text.

Kintsch, W. & Bowles, A. (2002). Metaphor comprehension: What makes a metaphor difficult to understand? *Metaphor and Symbol, 4,* 249-262.

Abstract: Comprehension difficulty was rated for metaphors of the form Noun 1 -is-a-Noun 2 ; in addition, participants completed frames of the form Noun 1 -is-_____ with their literal interpretation of the metaphor. Metaphor comprehension was simulated with a computational model based on Latent Semantic Analysis. The model matched participants' interpretations for both easy and difficult metaphors. When interpreting easy metaphors, both the participants and the model generated highly consistent responses. When interpreting difficult metaphors, both the participants and the model generated disparate responses.

Kintsch, W. (2002). The Potential of Latent Semantic Analysis for Machine Grading of Clinical Case Summaries. *Journal of Biomedical Informatics, 35,* 3-7.

Koper, R. & Tattersall, C. (2004). New directions for lifelong learning using network technologies. *British Journal of Educational Technology, 35,* 689-700.

Abstract: The requirements placed on learning technologies to support lifelong learning differ considerably from those placed on technologies to support particular fragments of a learning lifetime. The time scales involved in lifelong learning, together with its multi-institutional and episodic nature are not reflected in today's mainstream learning technologies and their associated architectures. The article presents an integrated model and architecture to serve as the basis for the realization of networked learning technologies serving the specific needs and characteristics of lifelong learners. The integrative model is called a "Learning Network" (LN) and its requirements and architecture are explored, together with the ways in which its application can help in reducing barriers to lifelong learning.

Notes: Pre-print in file

Koper, R., van Bruggen, J., Rusman, E., & Giesbers, B. (2005). *Learning Technology Development Programme - Positioning in learning networks*.

Koper, R., Rusman, E., & Sloep, P. (2005). Learning Network connecting people, organisations, software agents and learning resources to

establish the emergence of effective lifelong learning. *LLine: Lifelong Learning in Europe, 9,* 18-27.

Abstract: This article argues that the provision of lifelong learning opportunities needs to be based

on well-thought-through integrated models. These models should merge pedagogical, organisational and technological perspectives and meet requirements for the provision of lifelong learning opportunities. This article also claims that these requirements cannot be met by existing educational models and tools. The Learning Networks model is offered as an alternative, feasible model for ICT-network supported lifelong learning. A Learning Network is defined as an ensemble of actors, institutions and learning resources which are mutually connected through and supported by information and communication

technologies in such a way that the network self-organises and thus gives rise to effective lifelong learning.

Kurby, C. A., Wiemer-Hastings, K., Ganduri, N., Magliano, J. P., Mills, K. K., & McNamara, D. S. (2003). Computerizing reading training: Evaluation of a latent semantic analysis space for science text. *Behavior Research Methods, Instruments & Computers, 35,* 244-250.

Abstract: The effectiveness of a domain-specific latent semantic analysis (LSA) in assessing reading strategies was examined. Students were given self-explanation reading training (SERT) and asked to think aloud after each sentence in a science text. Novice and expert human raters and two LSA spaces (general reading, science) rated the similarity of each think-aloud protocol to benchmarks representing three different reading strategies (minimal, local, and global). The science LSA space correlated highly with human judgments, and more highly than did the general reading space. Also, cosines from science LSA spaces can distinguish between different levels of semantic similarity, but may have trouble in distinguishing local processing protocols. Thus, a domain-specific LSA space is advantageous regardless of the size of the space. The results are discussed in the context of applying the science LSA to a computer-based version of SERT that gives online feedback based on LSA cosines. (PsycINFO Database Record (c) 2003 APA, all rights reserved)

Laham, D. (1997). Latent Semantic Analysis approaches to categorization. In Proceedings of the 19th annual meeting of the Cognitive Science Society (pp. 979).

Laham, D., Bennett, W., & Landauer, T. K. (2000). An LSA-Based Software Tool for Matching Jobs, People, and Instruction. *Interactive Learning Environments, 8,* 171-185.

Abstract: New LSA-based agent software helps to identify required job knowledge, determine which members of the workforce have the knowledge, pinpoint needed retraining content, and maximize training and retraining efficiency. The LSA-based technology extracts semantic information about people, occupations, and task-experience contained in natural-text databases. The various kinds of information are all represented in the same way in a common semantic space. As a result, the system can match or compare any of these objects with any one or more of the others. To demonstrate and evaluate the system, we analyzed tasks and personnel in

three Air Force occupations. We measured the similarity of each airman to each task and estimated how well each airman could replace another. We also demonstrated the potential to match knowledge sub-components needed for new systems with ones contained in training materials and with those possessed by individual airmen. It appears that LSA can successfully characterize tasks, occupations and personnel and measure the overlap in content between instructional courses covering the full range of tasks performed in many different occupations. Such analyses may suggest where training for different occupations might be combined, where training is lacking, and identify components that may not be needed at all. In some instances it may suggest ways in which occupations might be reorganized to increase training efficiency, improve division of labor efficiencies, or redefine specialties to produce personnel capable of a wider set of tasks and easier reassignment.

Laham, D. (2001). *Automated content assessment of text using latent semantic analysis to simulate human cognition.* Univ Microfilms International, US.

Abstract: Latent Semantic Analysis (LSA) is both a theory of human knowledge representation and a method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text. The underlying idea is that the aggregate of all the word contexts in which a given word does and does not appear provides a set of mutual constraints that largely determines the similarity of meaning of words and sets of words to each other. Simulations of psycholinguistic phenomena show that LSA reflects similarities of human meaning effectively. The adequacy of LSA's reflection of human knowledge has been established in a variety of ways. For example, its scores overlap those of humans on standard vocabulary and subject matter tests; it mimics human word sorting and category judgments; it simulates word-word and passage-word lexical priming data; it accurately estimates learnability of passages by individual students and the quality and quantity of knowledge contained in an essay. To assess essay quality, LSA is first trained on domain-representative text. Then student essays are characterized by LSA representations of the meaning of their contained words and compared with essays of known quality on degree of conceptual relevance and amount of relevant content. Over many diverse topics, LSA scores agreed with human experts as accurately as expert scores agreed with each other. LSA has also been used to characterize tasks, occupations and personnel and measure the overlap in content between

instructional courses covering the full range of tasks performed in many different occupations. It extracts semantic information about people, occupations, and task-experience contained in natural-text databases. The various kinds of information are all represented in the same way in a common semantic space. As a result, the system can match or compare any of these objects with any one or more of the others. LSA-based agent software can help to identify required job knowledge, determine which members of the workforce have the knowledge, pinpoint needed retraining content, and maximize training and retraining efficiency. Computational models of concept relations using LSA representations demonstrate that categories can be emergent and self-organizing based exclusively on the way language is used in the corpus without explicit hand-coding of category membership or semantic features. LSA modeling also shows that the categories which are most often impaired in category specific semantic disnomias are those that show the most internal coherence in LSA representational structure. If brain structure corresponds to LSA structure, the identification of concepts belonging to strongly clustered categories should suffer more than weakly clustered concepts when their representations are partially damaged. (PsycINFO Database Record (c) 2002 APA, all rights reserved)

Laham, D., Bennett, W., & Derr, M. (2002). Latent Semantic Analysis for Career Field Analysis and Information Operations. [On-line]. Available: http://www.k-a-t.com/papers/ab-careerField2002.shtml

Abstract: This paper reviews two current Air Force Research Laboratory / Human Effectiveness Directorate (AFRL/HEA) efforts that are maturing Latent Semantic Analysis (LSA) tools for the Air Force. The first effort is developing new LSA-based agent software that helps decision makers to identify required job knowledge, determine which members of the workforce have the knowledge, pinpoint needed retraining content, and maximize training and retraining efficiency. Modern organizations are increasingly faced with rapid changes in technology and missions and need constantly changing mixes of competencies and skills. Assembling personnel with the right knowledge and experience for a task is especially difficult when there are few experts, unfamiliar devices, redefined goals, and short lead-times for training and deployment. LSA is being used to analyze course content and materials from current training pipelines and to identify appropriate places in alternative structures where that content can be reused. This saves time for training developers since the preexisting content has already been validated as a part of its earlier

application.

AFRL/HEA's second research effort involves a demonstration of a combined speech-to-text and LSA-based software agent for embedding automatic, continuous, and cumulative analysis of verbal interactions in individual and team operational environments. The agent will systematically parse and evaluate verbal communication to identify critical information and content required of many of today's AF operators. LSA is promising new technology that has significant potential for assisting operators in the performance of their tasks because it can "listen" and in almost real-time evaluate free-form verbal communication from a variety of sources and match content to stored language dictionaries. One application of this technology being explored is tracking and scoring the tactical communications that occur between the members of a four-ship air combat flight and their weapons director to identify areas of training need and as an additional tool for assessing the efficacy of DMT scenarios and missions.

Landauer, Laham, D., and Foltz, Peter W. (2000). The intelligent essay assessor. Putting knowledge to the test. *IEEE Intelligent Systems,* 27-31.

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the 19th Annual Conference of the Cognitive Science Society* (pp. 412-417). Mahwah, NJ: Lawrence Erlbaum Associates.

Notes: Doet verslag van twee experimenten waarin beoordeling van essays door LSA en menselijke beoordelaars (ETS deskundigen) worden vergeleken. **Experiment 1:** 94 undergraduates vervaardigden een tekst van ongeveer 250 woorden over anatomie, functie en doel van het menselijk hart. Twee deskundige beoordelaars namen kennis van relevant bronmateriaal en overlegden welke inhoud in de essays voor moest komen. Vervolgens beoordeelden ze de essays op een vijfpuntsschaal. De studenten kregen een 40 punten toets met korte antwoord vragen.LSA werd getraind met 27 artikelen uit de Grolier's Academic American Encyclopedia: 94 dimensionale ruimte voor 830 zinnen en 3034 unieke woorden met filtering van een stoplist. Voor ieder essay werd een vector berekend door het gemiddelde te nemen van de vectoren van de woorden in het essay. Er werd op twee manieren gescored. Methode 1 - twee componenten: a) bereken de cosinus van de vector voor het essay en alle andere essays. De score die het essay kreeg werd gebaseerd op de tien meest

overeenkomende essays: een gewogen gemiddelde (cosinus als weging) van de gemiddelde score van de beoordelaars en b) vectorlengte van het essay. Methode 2: geen gebruik van menselijke beoordelaars, maar vergelijking met een tekst geschreven door een expert (uit een inleidend biologie werk). Resultaten: menselijke beoordelaars: r = .77. LSA - beoordelaars + gemiddelde beoordelaars: r=.68 , r=.77, r=.77. Correlaties met toetsscore: gemiddelde beoordelaar - toets: r=.70, LSA r+.81. Methode 2 (gouden standaard): LSA - beoordelaars r+.64, .71, .72. LSA en toetsscore: r=.77. Bij methode 2 werd ook vectorlengte nader bekeken: alleen de vectoren waarin de technische begrippen zitten dragen significant bij aan de voorspelling. (correlatie vectorlengte met alle woorden - toets r=.77; correlatie voor vector met alleen technische begrippen r=.72) NB dit is dus nog beter dan de menselijke beoordelaars.

Experiment 2 is in feite een replicatie maar nu met 273 Inleiding Psychologie studenten die in tien minuten een essay schreven over operant conditioneren, hechting bij kinderen of afasie. Twee inhoudsdeskundigen als beoordelaars. De beoordelaars haalden gemiddeld over alle essays een correlatie van .65, maar tussen de essays deden zich grote verschillen voor (hechting r=.19). Methode 1 voor beoordeling is kwetsbaar voor geringe interraterbetrouwbaarheid. De correlaties liggen hier dan ook lager: LSA - gemiddelde oordeel r=.64 en loopt voor de drie topics uiteen van .61 tot .71.

Landauer, T. K., Laham, D., & Foltz, P. W. (1998). Learning human-like knowledge by Singular Value Decomposition: A progress report. *Advances in Neural Information Processing Systems, 10,* 45-51.

Abstract: Singular value decomposition (SVD) can be viewed as a method for unsupervised training of a network that associates two classes of events reciprocally by linear connections through a single hidden layer. SVD was used to learn and represent relations among very large numbers of words (20k-60k) and very large numbers of natural text passages (1k-70k) in which they occurred. The result was 100-350 dimensional "semantic spaces" in which any trained or newly added word or passage could be represented as a vector, and similarities were measured by the cosine of the contained angle between vectors. Good accuracy in simulating human judgments and behaviors has been demonstrated by performance on multiple-choice vocabulary and domain knowledge tests, emulation of expert essay evaluations, and in several other ways. Examples are also given of how the kind of knowledge extracted by this method can be applied.

Landauer, T. K. (2002). Applications of Latent Semantic Analysis. In *Proceedings of the 24th Annual Meeting of the Cognitive Science Society*.

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211-240.

Abstract: How do people know as much as they do with as little information as they get? The problem takes many forms; learning vocabulary from text is an especially dramatic and convenient case for research. A new general theory of acquired similarity and knowledge representation, latent semantic analysis (LSA), is presented and used to successfully simulate such learning and several other psycholinguistic phenomena. By inducing global knowledge indirectly from local co-occurrence data in a large body of representative text, LSA acquired knowledge about the full vocabulary of English at a comparable rate to schoolchildren. LSA uses no prior linguistic or perceptual similarity knowledge; it is based solely on a general mathematical learning method that achieves powerful inductive effects by extracting the right number of dimensions (e.g., 300) to represent objects and contexts. Relations to other theories, phenomena and problems are sketched. (PsycINFO Database Record (c) 2002 APA, all rights reserved)

Landauer, T. K. (1998). Learning and representing verbal meaning: The Latent Semantic Analysis Theory. *Current Directions in Psychological Science, 7,* 161-164.

Abstract: Latent semantic analysis (LSA) is a theory of how word meaning--and possibly other knowledge--is derived from statistics and experience, and of how passage meaning is represented by combinations of words. Given a large and representative sample of text, LSA combines the way thousands of words are used in thousands of contexts to map a point for each into a common semantic space. LSA goes beyond pair-wise co-occurrence or correlation to find latent dimensions of meaning that best relate every word and passage to every other. After learning from comparable bodies of text, LSA has scored almost as well as humans on vocabulary and subject-matter tests, accurately simulated many aspects of human judgment and behavior based on verbal meaning, and has been successfully applied to measure the coherence and conceptual content of text. The surprising success of LSA has implications for the nature of generalization and language. (PsycINFO Database Record (c) 2002 APA, all rights reserved)

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25,* 259-284.

Abstract: Latent Semantic Analysis (LSA) is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text (T. K. Landauer & S. T. Dumais, 1997). The underlying idea is that the aggregate of all the word contexts in which a given word does and does not appear provides a set of mutual constraints that largely determines the similarity of meaning of words and sets of words to each other. The adequacy of LSA's reflection of human knowledge has been established in a variety of ways. For example, its scores overlap those of humans on standard vocabulary and subject matter tests; it mimics human word sorting and category judgments; it simulates word-word and passage-word lexical priming data; and it accurately estimates passage coherence, learnability of passages by individual students, and the quality and quantity of knowledge contained in an essay. (PsycINFO Database Record (c) 2002 APA, all rights reserved)

Landauer, T. K. (1999). Latent semantic analysis: A theory of the psychology of language and mind. *Discourse Processes, 27,* 303-310.

Abstract: Comments on a recent article by T. K. Landauer, P. W. Foltz, and D. Laham (see record 1998-10451-004), in which the authors described a computational model called Latent Semantic Analysis (LSA) and summarized its successful simulations of a variety of human performance phenomena that depend on word and passage meaning. Subsequent articles in the same special issue of "Discourse Processes" reported details of several of these studies. Charles Perfetti (1998), in a commentary, agreed that LSA is a useful research tool, but argued that it should not be regarded as a plausible theory of mind because it is based on learning from co-occurrence data. In his response, the author shows why this objection lacks merit, and clarifies what LSA has to offer, suggesting that LSA does not handle all aspects of language processing, but offers a biologically and psychologically plausible mechanistic explanation of the acquisition, induction, and representation of verbal meaning. (PsycINFO Database Record (c) 2002 APA, all rights reserved)

Landauer, T. K. & Psotka, J. (2000). Simulating Text Understanding for Educational Applications with Latent Semantic Analysis: Introduction to LSA. *Interactive Learning Environments, 8,* 73-86.

Landauer, T. K. (2001). Single representations of multiple meanings in latent semantic analysis. In D.S.Gorfein (Ed.), *On the consequences of meaning selection: Perspectives on resolving lexical ambiguity* (pp. 217-232). Washington, DC, US: American Psychological Association.

Abstract: Latent Semantic Analysis (LSA) is a psychological model and computational simulation intended to mimic and help explain the way that humans learn and represent the meaning of words, text, and other knowledge. In this chapter, I briefly describe the underlying theoretical and computational machinery of LSA, review some of the surprising things it is able to do, and discuss some of its limitations and possibilities for future development. I concentrate on what LSA has to say about multiple word meanings, where it succeeds and fails, and what is needed to fix it. For researchers and theorists concerned with word meanings and ambiguity, the most important implication of the LSA theory is that it questions the idea that different senses of a word have separate and discrete representations that are individually disambiguated. Instead, it represents a word meaning as a single point in a very high-dimensional semantic space. In LSA, the acquisition of a word meaning is an irreversible mathematical melding of the meanings of all the contexts in which it has been encountered. In comprehension, words are not disambiguated by sense one at a time. (PsycINFO Database Record (c) 2002 APA, all rights reserved)

Landauer, T. K. (2002). On the computational basis of learning and cognition: Arguments from LSA. In B.H.Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory, Vol. 41* (pp. 43-84). San Diego, CA, US: Academic Press.

Abstract: The theme of this chapter is the relation between the nature of evolutionarily determined computational processes that support learning. Examples of this focus are neural mechanism conjectures, connectionist modeling, and mathematical learnability theory. The discussion is organized in a somewhat spiral fashion. Landauer first raises the issue of the basic elements of empirical association and illustrates it by the case of learned object recognition. This leads to the hypothesis that the choice of optimal elements may provide a small part of the solution of the problem; what is done with the co-occurring elements appears to be more important. He then moves to the learning of word and passage meaning because this domain exhibits the problem in a manner that is convenient to model; we can give a computer the very same mass of perceptual input that literate humans use for much of their learning. Landauer first shows how a different kind of co-occurrence data and a different form of computation can yield

more knowledge that has usually been supposed. Next, he explains how these ideas are implemented in the latent semantic analysis (LSA) learning model through singular value decomposition. He then lists a variety of human verbal comprehension performances that LSA simulates well. (PsycINFO Database Record (c) 2003 APA, all rights reserved)

Langer, H., Lüngen, H., & Bayerl, P. S. (29-5-2004). Towards Automatic Annotation of Text Type Structure. XBrac-Workshop. Lisbon, Portugal. 12-5-2005.

Ref Type: Unpub. contribution symposium

Abstract: Experiments Using an XML-Annotated Corpus and Automatic Text Classification Methods

Lemaire, B. (2006). Readings in Latent Semantic Analysis for Cognitive Science and Education. [On-line].

Available: http://www-leibniz.imag.fr/perso/s1/blemaire/public_html/lsa.html

Notes: Bibliography

Lemaire, B. & Dessus, P. (2001). A System to Assess the Semantic Content of Student Essays. *Journal of Educational Computing Research, 24,* 305-320.

Notes: ABSTRACT This paper presents Apex, a system that can automatically assess a student essay based on its content. It relies on Latent Semantic Analysis, a tool which is used to represent the meaning of words as vectors in a high-dimensional space. By comparing an essay and the text of a given course on a semantic basis, our system can measure how well the essay matches the text. Various assessments are presented to the student regarding the topic, the outline and the coherence of the essay. Our experiments yield promising results.

Lemaire, B. & Denhière, G. (2004). Cognitive Models based on Latent Semantic Analysis. ICCM'2003.

Ref Type: Unpub. contribution symposium

Letsche, T. A. (1997). Large-Scale Information Retrieval with Latent Semantic Indexing. *Informatics and Computer Science, 100,* 105-137.

Lloyd, R. & Shakiban, C. (2004). Improvemnets in Latent Semantic Analysis. *American Journal of Undergraduate Research, 3*.

Abstract: This paper proposes and examines modifications for the method of Latent Semantic Analysis (LSA). Several new local and global weight functions, along with normalization routines, are disclosed. Changes in the general structure of LSA are discussed. An application of LSA, in

which the method is used to filter advertisements in e-mail, proves the worthiness of the advancements.

Magliano, J. P., Wiemer-Hastings, K., Millis, K. K., Munoz, B. D., & McNamara, D. (2002). Using latent semantic analysis to assess reader strategies. *Behavior Research Methods, Instruments & Computers, 34,* 181-188.

Abstract: Tested a computer-based procedure for assessing reader strategies that was based on verbal protocols that utilized latent semantic analysis (LSA). 212 Students were given self-explanation-reading training (SERT), which teaches strategies that facilitate self-explanation during reading, such as elaboration based on world knowledge and bridging between text sentences. During a computerized version of SERT practice, students read texts and typed self-explanations into a computer after each sentence. The use of SERT strategies during this practice was assessed by determining the extent to which students used the information in the current sentence versus the prior text or world knowledge in their self-explanations. This assessment was made on the basis of human judgments and LSA. Both human judgments and LSA were remarkably similar and indicated that students who were not complying with SERT tended to paraphrase the text sentences, whereas students who were compliant with SERT tended to explain the sentences in terms of what they knew about the world and of information provided in the prior text context. The similarity between human judgments and LSA indicates that LSA will be useful in accounting for reading strategies in a Web-based version of SERT. (PsycINFO Database Record (c) 2002 APA, all rights reserved)

Magliano, J. P. & Millis, K. K. (2003). Assessing reading skill with a think-aloud procedure and latent semantic analysis. *Cognition & Instruction, 21,* 251-283.

Abstract: The viability of assessing reading strategies is studied based on think-aloud protocols combined with Latent Semantic Analysis (LSA). Readers in two studies thought aloud after reading specific focal sentences embedded in two stories. LSA was used to estimate the semantic similarity between readers' think-aloud protocols to the focal sentences and sentences in the stories that provided direct causal antecedents to the focal sentences. Study 1 demonstrated that according to human- and LSA-based assessments of the protocols, the responses of less-skilled readers semantically overlapped more with the focal sentences than with the causal antecedent sentences, whereas the responses of skilled readers overlapped with

these sentences equally. In addition, the extent that the semantic overlap with causal antecedents was greater than the overlap with the focal sentences predicted performance on comprehension test questions and the Nelson-Denny test of reading skill. Study 2 replicated these findings and also demonstrated that the semantic overlap scores (based on the protocols) predicted recall for stories that were read silently. Together, the findings supported the viability of developing a computerized assessment tool using verbal protocols and LSA. (PsycINFO Database Record (c) 2003 APA, all rights reserved)

Manning, C. D. & Schütze, H. (2004). Contents of Foundations of Statistical Natural Language Processing. [On-line]. Available: http://nlp.stanford.edu/fsnlp/promo/

Notes: Web page with contents of the book, links to sample chapters and reviews

Marcu, D. (2000). *The theory and practice of discourse parsing and summarization*. Cambridge, MA: Bradford / MIT Press.

McCauley, L. (2002). Using Latent Semantic Analysis tot aid speech recognition and understanding.

Ref Type: Unpublished Work

Abstract: Generally, speech recognition engines can employ two different grammar methods, rule and dictation, to recognize an utterance. The purpose of these grammars is to constrain the search space in a way that anticipates the speaker's utterance. The research described in this paper attempts to maintain the accuracy of a rule grammar without limiting the speaker to rigorous phraseology. Latent Semantic Analysis (LSA) is used to connect specific grammar rules with the meanings underlying matching phrases resulting in utterances being matched to knowledge base elements even though the exact phrase did not match any grammar rule. A separate knowledge base is used to dynamically add or remove grammar rules in the speech recognition engine as the conversation context changes. Finally, a learning technique is used to create new regular expressions based on utterances that matched semantically through LSA..

Microsoft (2005). Encarta 2005 Standard [Computer software].

Miller, T. (2003). Essay assessment with latent semantic analysis. *Journal of Educational Computing Research, 29,* 495-512.

Abstract: Latent semantic analysis (LSA) is an automated, statistical technique for comparing the semantic similarity of words or documents. In this paper, I examine the application of LSA to

automated essay scoring. I compare LSA methods to earlier statistical methods for assessing essay quality, and critically review contemporary essay-scoring systems built on LSA, including the Intelligent Essay Assessor, Summary Street, State the Essence, Apex, and Select-a-Kibitzer. Finally, I discuss current avenues of research, including LSA's application to computer-measured readability assessment and to automatic summarization of student essays.

Moertl, P. M. (2003). *Elicitation of knowledge differences in reading comprehension using latent semantic analysis with multiple semantic spaces.* Univ Microfilms International, US.

Abstract: Previous research has proposed Latent Semantic Analysis (LSA) as a model and technique of knowledge representation that represents knowledge differences in single semantic spaces (e.g. Grolier's Academic American Encyclopedia, Landauer & Dumais 1997). In this project, LSA knowledge representations were constructed in multiple semantic spaces to represent user knowledge differences for adaptive information retrieval. Semantic spaces with varying degrees of background knowledge were constructed for two versions of a story that participants had read. The two versions induced either complete or incomplete story comprehension. The results indicated that optimal LSA representations depended on the level of story comprehension: LSA representations that were derived from semantic spaces of any size resembled participants' complete story comprehension but matched incomplete story comprehension only if semantic spaces included sufficient information. Larger semantic spaces captured more background knowledge than smaller spaces (Experiment 2). This led to the conclusion that participants with incomplete comprehension relied more on background knowledge to rate word pair relatedness than in the Solved condition where they relied more on story knowledge. Comparing LSA representations in multiple semantic spaces was found to be a viable means for representing knowledge dependent on a reader's background. Implications of these findings for the representation of user knowledge for automated adaptive information retrieval are discussed. (PsycINFO Database Record (c) 2003 APA, all rights reserved)

Moscoso del Prado Martin, F. & Sahlgren, M. (2002). An integration of Vector-Based Semantic Analysis and Simple Recurrent Networks for the automatic acquisition of lexical representations from unlabeled corpora. In *Proceedings of the Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data Workshop*.

Nakov, P. (2000). Getting better results with Latent Semantic Indexing. In *Proceedings of the students presentations at the European Summer School in Logic Language and Information (ESSLLI'00)* (pp. 156-166).

Nakov, P., Popova, A., & Mateev, P. (2001). Weight functions impact on LSA performance. In Tzigov Chark (Ed.), *Recent Advances in Natural language processing (RANLP'2001).* (pp. 187-193).

Nakov, P., Valchanova, E., & Angelova, G. (2003). Towards deeper understanding of the LSA Performance. In *Recent Advances in Natural Language processing (RANLP'2003)* (pp. 311-318).

Olde, B. A., Franceschetti, D. R., Karnavat, A., Graeser A.C., & Tutoring Research Group (2002). The Right Stuff: Do You Need to Sanitize Your Corpus When Using Latent Semantic Analysis. In *24th Annual Meeting of the Cognitive Science Society* (pp. 708-713). Fairfax.

Abstract: Student responses to conceptual physics questions were analyzed with latent semantic analysis (LSA), using different text corpora. Expert evaluations of student answers to questions were correlated with LSA metrics of the similarity between student responses and ideal answers. We compared the adequacy of several text corpora in LSA performance evaluation, including the inclusion of written incorrect reasoning and tangentially relevant historical information. The results revealed that there is no benefit in meticulously eliminating the wrong or irrelevant information that normally accompanies a textbook. Results are also reported on the impact of corpus size and the addition of information that is not topic relevant.

Page, E. (1966). The imminence of grading essays by computer. *Phi Delta Kappan,* 238-243.

Perfetti, C. (1998). The Limits of Co-Occurence. *Discourse Processes, 25,* 363-377.

Picciano, A. (2004). *Educational Research Primer*. London: Continuum.

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

Quesada, J. F., Kintsch, W., & Gomez, E. (2001). A Computational Theory of Complex Problem Solving Using the Vector Space Model (part I): Latent Semantic Analysis, Through the Path of Thousands of Ants. *Cognitive research with Microworlds, 43,* 117-131.

Abstract: For years, researchers have argued that Complex Problem Solving (CPS) is plagued with methodological problems. The interest of this research paradigm, a hybrid between field studies and experimental ones, is tied to the success of methodological advances that enable

performance to be analyzed. This paper introduces a new, abstract conceptualization of *microworlds* research based on two theoretical lines: (1) a representational problem, where protocols can be seen as objects in a feature space and, (2) a similarity measure problem, where a similarity metric has to be proposed. To materialize this conceptualization we introduce Latent Semantic Analysis (LSA), a machine-learning model that induces representations of the meaning of words by analyzing the relation between words and passages in large bodies of representative text, and describe how LSA can be implemented as a theory and technique to analyze performance in CPS, using actions or states as units instead of words and trials instead of text passages. Basic examples of application are provided, and advantages and disadvantages are discussed.

Quesada, J. F., Kintsch, W., & Gomez, E. (2001). A Computational Theory of Complex Problem Solving Using the Vector Space Model (part II): Latent Semantic Analysis Applied to Empirical Results from Adaptation Experiments. *Cognitive resarch with Microworlds, 43,* 117-131.

Abstract: The literature of complex problem solving and system control has focused on how to improve the adaptation of operators to new, unpredictable circumstances. The present work reviews the main methodologies and assumptions that are currently being used in complex, dynamic task to answer questions regarding the adaptability problem, i.e. the work on DuressII (Vicente and Collaborators) and on Firechief (Cañas and collaborators). Some methodological problems for Cañas et al. analysis assumptions that could have important consequences in the results obtained are discussed. This study proposes Latent Semantic Analysis (LSA) as an alternative that remedies some of the flaws and adds some interesting new possibilities of analysis, such as coherence measures to assess performance changes in a similar vein as the Within-trial Trajectory Deviation (WTD) used in continuous systems such as DuressII. The study uses an LSA corpus created from the experimental data generated by past experiments in Firechief on adaptation to unpredictable task changes to replicate and extend the results previously obtained. The new LSA approach and results obtained are discussed. The fact that results from both microworlds could be explained by LSA with no modifications in its basic assumptions promises a future common theory and method of complex problem solving.

Quesada, J. F., Kintsch, W., & Gomez, E. (2002). A Computational Theory of Complex Problem Solving using Latent Semantic Analysis. In W. D. &. S. Gray (Ed.), *24th Annual Conference of the*

Cognitive Science Society (pp. 750-755). Fairfax, VA. Lawrence Erlbaum Associates, Mahwah, NJ.

Abstract: Complex Problem Solving (CPS) is a hybrid between field studies and experimental studies. This paper introduces a new, abstract conceptualization of *microworlds* research based on two innovations: (1) a problem representation, which treats protocols as objects in a feature space and, (2) a similarity metric which is defined in this problem space. Latent Semantic Analysis (LSA) is used to analyze performance in CPS, using actions or states as units instead of words and trials instead of text passages. Basic examples of applications are provided, and advantages and limitations are discussed.

Quesada, J. F. (2003). Introduction to Latent Semantic Analysis and Latent Problem Solving Analysis. In *Latent Problem Solving Analysis (LPSA): A computational theory of representation in complex, dynamic problem solving tasks* (pp. 22-35).

Quesada, J. F., Kintsch, W., & Gomez, E. (2003). Automatic Landing Technique Assessment using Latent Problem Solving Analysis. In R. Alterman & D. Kirsch (Eds.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society* Boston, MA.

Abstract: Latent Problem Solving Analysis is applied to model the decision processes of expert instructors judging professional pilots' landing technique in a B747 flying simulator, showing that that a memory-based model can do well in the absence of more conscious, logical processes.

Rehder, B., Schreiner, M. E., Wolfe, M. B., Laham, D., Landauer, T. K., & Kintsch, W. (1998). Using latent semantic analysis to assess knowledge: some technical considerations. *Discourse Processes, 25,* 337-354.

Notes: In dit artikel onderzoeken Rehder et al. welke alternatieven er zijn voor de gebruikte LSA-technieken in de analyse van Wolfe et. al. In de eerste plaats onderzoeken ze wat het effect is van het weglaten van alle niet-technische termen uit de studentessays. Interessant genoeg bleken zowel de technische als niet-technische termen behoorlijk hoog te correleren met de oorspronkelijke cos tussen de vectoren (.94 en .86. Ze dragen beide significant bij aan de voorspelling van de voorkennismeting. In een tweede onderzoek werd het effect onderzocht van het aantal woorden in het essay op de voorspelling (cos) van voormeting. Een minimum van 70 lijkt daarbij noodzakelijk en boven de 200 woorden neemt de meeropbrenst snel af. Tenslotte werd onderzoek gedaan naar het effect van een reeks andere vectorgebaseerde maten zoals

Euclidische afstand tussen de vectoren en de lengte van het essayvector (NB de gebruikelijke cos is gebaseerd op orthonormale vectoren!). Voorzichtige conclusie is dat alleen de lengte van de E-vector een significante bijdrage aan de voorspelling levert (cf PEG en e-rater). De lengte van deze vector is vooral afhankelijk van de technische termen in het essay, maar ook van meer algemene kennis (encyclopedie bijvoorbeeld). Ook in niet gepubliceerd onderzoek van Laham en Landauer bleek vectorlengte een significante voorspeller.Interessant is de laatste vraag die aan de orde komt: wat te doen als er groepen ppn zijn die duidelijk verschillen in hun voorkennis, bijvoorbeeld experts en beginners? Beide zullen, in de cosinus-maten, aanzienlijk afwijken van de brontekst, maar om geheel verschillende redenen. Rehder et al. hebben gewerkt met verschillende multidimensionale schaalmethoden en konden zo laten zien welke plaats de experts en beginners in de semantische ruimte innamen. (JBR: dit onderzoek schijnt geen follow-up te hebben gehad).

Rosario, B. (2000). *Latent Semantic Indexing: An Overview* (Rep. No. INFOSYS 240).

Shapiro, A. M. & McNamara, D. S. (2000). The use of latent semantic analysis as a tool for the quantitative assessment of understanding and knowledge. *Journal of Educational Computing Research, 22,* 1-36.

Abstract: Investigated whether a latent semantic analysis (LSA), a statistical model of word usage, more accurately reflects the factual or conceptual knowledge contained in written material. 60 college students participated. Exp 1 compared LSA analyses of essays to human-generated scores. It also compared the LSA output to several measures of conceptual structure. Exp 2 correlated LSA analyses of transcribed recall protocols with a series of comprehension measures that were designed to vary in the degree to which they reflect conceptual or factual knowledge. Result show that LSA analyses were a stronger reflection of the text-based knowledge represented by essays and recall protocols than conceptual knowledge. Findings indicate that LSA performed best when trained in a content area specific to the material to be analyzed. The results are discussed with respect to the application of LSA analyses in the classroom and laboratory. (PsycINFO Database Record (c) 2002 APA, all rights reserved)

Soto, R. (1998). Learning and performing by Exploration: Label quality measured by Latent Semantic Analysis.

Ref Type: Unpublished Work

Abstract: Models of learning and performing by exploration assume that the *semantic distance* between task descriptions and screen labels controls in part the usersÕ search strategies. Nevertheless, none of the models has an objective way to compute semantic distance. In this study, participants performed twelve tasks by exploration and were tested for recall after a 1-week delay. Latent Semantic Analysis was used to compute the semantic similarity between the task descriptions and the labels in the applicationÕs menu system. When the labels were close in the semantic space to the task descriptions, subjects performed the tasks faster. LSA could be incorporated into any of the current models, and it could be used to automate the evaluation of computer applications for ease of learning and performing by exploration.

Stahl, G. (1997). Allowing learners to be articulate: incorporating automated text evaluation into collaborative software environments. Proposal to the McDonnell foundation.

Ref Type: Generic

Steinhart, D. (2001). *Summary Street: an intelligent tutoring system for improving student writing through the use of latent semantic analysis.* Institute of Cognitive Science, University of Colorado. Boulder.

Abstract: This dissertation describes the design, evolution, and testing of *Summary Stree*t, an intelligent tutoring system which uses Latent Semantic Analysis (LSA) to support writing and revision activities. *Summary Street* provides various kinds of automatic feedback, primarily whether a summary adequately covers important source content and fulfills other requirements, such as length. The feedback allows students to engage in extensive, independent practice in writing and revising without placing excessive demands on teachers for feedback. The efficacy of this system was examined in three classroom studies in a Boulder County school.

In the first study, students read texts about three Mesoamerican civilizations and then composed summaries of those texts. One summary was produced using Summary Street, while the other two were produced using a traditional word processor. The students who used Summary Street to

summarize the most difficult text produced better summaries than those students who used a word processor.

In the second study, students learned about the human circulatory system and summarized two texts about the heart and lungs. The same pattern from the first study was observed-namely, the

students who summarized the more difficult lung text using Summary Street produced better summaries than those students who used a word processor. The third and final study produced the clearest results. Ten different texts were used in the study, and there was a correlation between text difficulty and the value of the feedback from Summary Street—the more difficult the text, the more Summary Street helped students write better summaries.
In addition to the aforementioned results, individual differences and issues regarding transfer are discussed. Finally, intelligent tutors and the role of technology in the classroom are examined, and Summary Street is compared to existing intelligent tutors.

Steyvers, M. & Tenenbaum, J. B. (2005). The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Sematic Growth. *Cognitive Science, 29,* 41-78.

Strait M.J., Haynes, J. A., & Foltz, P. W. (2000). Applications of Latent Semantic Analysis to Lessons Learned Systems. In D. W. Aha & R. Weber (Eds.), Menlo Park, CA: AAAI Press.
Abstract: This paper will present several examples of the application of Latent Semantic Analysis (LSA) to practical problems of knowledge acquisition, management and assessment. The purpose of this presentation is to make other knowledge management (KM) and artificial intelligence (AI) researchers aware of the value of LSA as an automated technique for improving the utility of Lessons Learned (LL) and similar knowledge and information management systems.

Takayama, Y., Flournoy, R. S., & Kaufmann, S. (1998). Information Mapping. Concept-based Information Retrieval based on Word Associations.
Ref Type: Unpub. contribution symposium

Turney, P. D., Litmann, M. L., Bigham, J., & Shnayder, V. (2003). Combining Independent Modules to Solve Multiple-Choice Synonym and Analogy Problems. In G. Angelova, K. Botcheva, R. Mitkov, & N. Nicolov (Eds.), *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-03)* (pp. 482-389).
Abstract: Existing statistical approaches to natural language problems are very coarse approximations to the true complexity of language processing. As such, no single technique will be best for all problem instances. Many researchers are examining ensemble methods that combine the output of successful, separately developed mod- ules to create more accurate

solutions. This pa- per examines three merging rules for combining probability distributions: the well known mixture rule, the logarithmic rule, and a novel product rule. These rules were applied with state-of-the-art results to two problems commonly used to assess human mastery of lexical semantics|synonym questions and analogy questions. All three merging rules result in ensembles that are more accurate than any of their component modules. The dierences among the three rules are not statistically signicant, but it is suggestive that the popular mixture rule is not the best rule for either of the two problems.

Turney, P. D., Litmann, M. L., Bigham, J., & Shnayder, V. (2004). Combining Independent Modules in Lexical Multiple-Choice Problems. In N. Nicolov, K. Botcheva, G. Angelova, & R. Mitkov (Eds.), *Recent Advances in Natural Language Processing III: Selected Papers from RANLP 2003* (pp. 101-110). John Benjamins.

Abstract: Existing statistical approaches to natural language problems are very coarse approximations to the true complexity of language processing. As such, no single technique will be best for all problem instances. Many researchers are examining ensemble methods that combine the output of multiple modules to create more accurate solutions. This paper examines three merging rules for combining probability distributions: the familiar mixture rule, the logarithmic rule, and a novel product rule. These rules were applied with state-of-the-art results to two problems used to assess human mastery of lexical semantics -- synonym questions and analogy questions. All three merging rules result in ensembles that are more accurate than any of their component modules. The differences among the three rules are not statistically significant, but it is suggestive that the popular mixture rule is not the best rule for either of the two problems.

van Bruggen, J. (2002). *Computerondersteund beoordelen van essays* (Rep. No. OTEC 2002/1). Heerlen: Onderwijstechnologisch expertisecentrum, Open Universiteit Nederland.

van Bruggen, J., Sloep, P., van Rosmalen, P., Brouns, F., Vogten, H., Koper, R. et al. (2004). Latent semantic analysis as a tool for learner positioning in learning networks for lifelong learning. *British Journal of Educational Technology, 35,* 729-738.

Abstract: As we move towards distributed, self-organized learning networks for lifelong learning to which multiple providers contribute content, there is a need to develop *new* techniques to determine where learners can be positioned in these networks. Positioning requires us to map characteristics of the learner onto characteristics of learning materials and curricula. Considering

the nature of the network envisaged, maintaining data on these characteristics and ensuring their integrity are difficult tasks. In this article we review the usability of Latent Semantic Analysis (LSA) to generate a common semantic framework for characteristics of the learner, learning materials and curricula. Although LSA is a promising technique, we identify several research topics that must be addressed before it can be used for learner positioning.

Wade-Stein, D. & Kintsch, E. (2003). *Summary Street: Interactive Computer Support for Writing* (Rep. No. 03-01(2003)). Colorado: University of Colorado.

Abstract: *Summary Street* is educational software that incorporates cognitive research on the development of summarization skills with the meaning representation method, called Latent Semantic Analysis. *Summary Street* provides students automatic feedback on the content of their summaries. The feedback is presented in an easy-to-grasp, graphic display that motivates students to improve their writing across multiple cycles of writing and revision on their own before handing it in to the teacher for final evaluation. The software thus has the potential to provide students with extensive writing practice without increasing the teacher's workload. In classroom trials sixth-grade students not only wrote better summaries when receiving content-based feedback from *Summary Stree*t, but also spent more than twice as long engaged in the writing task. Improvement in content scores was greater when students were summarizing more difficult texts. The authors suggest that *Summary Street* could be adapted to a wide variety of instructional goals beyond summary writing.

Wall, M. E., Rechtsteiner, A., & Rocha, L. M. (2003). Singular value decomposition and principal component analysis. In D.P.Berrar, W. Dubitzky, & M. Granzow (Eds.), *A practical apporach to microarray data analysis.* (pp. 91-109). Norwell, MA.: Kluwer.

Whittington, D. & Hunt, H. (1999). Approaches to the computerized assessment of free text responses. In *Proceedings of the 3rd CAA Conference*.

Wiemer-Hastings, P. (1999). How latent is Latent Semantic Analysis? In *Proceedings of the Sixteenth International Joint Congress on Artificial Intelligence* (pp. 932-937). San Francisco: Morgan Kaufmann.

Wiemer-Hastings, P., Wiemer-Hastings, K., & Graesser, A. C. (1999). Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis. In S. P. Lajoie & Vivet M. (Eds.),

Artificial Intelligence in Education (pp. 535-542). Amsterdam: IOS Press.

Abstract: AutoTutor is an intelligent tutor that interacts smoothly with the student using natural language dialogue. This type of interaction allows us to extend the domains of tutoring. We are no longer restricted to areas like mathematics and science where interaction with the student can be limited to typing in numbers or selecting possibilities with a button. Others have tried to implement tutors that interact via natural language in the past, but because of the diculty of understanding language in a wide domain, their best results came when they limited student answers to single words. Our research directly addresses the problem of understanding what the student naturally says. One solution to this problem that has recently emerged is Latent Semantic Analysis (LSA). LSA is a statistical, corpus-based natural language understanding technique that supports similarity comparisons between texts. The success of this technique has been described elsewhere [3, 5, for example]. In this paper, we give an overview of LSA and how it is used in our tutoring system. Then we focus on an important issue for this type of corpus-based natural language analysis, namely, how large must the training corpus be to achieve ecient performance? This paper describes two studies which address this question, and systematically tests the kind of texts needed in the corpus. We discuss the implications of these results for tutoring systems in general.

Wiemer-Hastings, P. & Graesser, A. C. (2000). Select-a-Kibitzer: a computer tool that gives meaningful feedback on student compositions. *Interactive Learning Environments, 8,* 149-169.

Wiemer-Hastings, P. & Zipitria, I. (2001). Rules for syntax, vectors for semantics. In *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society* (pp. 1112-1117). Mahwah,NJ: Lawrence Erlbaum Associates.

Notes: Wiemer-Hastings, P. & Zipitria, I. (2001). Rules for Syntax, Vectors for Semantics. In Proceedings of the 23rd Annual Conference of the Cognitive Science Society, (pp. 1112-1117). Mahwah, NJ: Lawrence Erlbaum Associates.

Wiemer-Hastings, P. (2001). List of Publications. [On-line]. Available: http://www.cogsci.ed.ac.uk/~peterwh/papers/

Notes: last update 29 may, 2001

Wiemer-Hastings, P. (2002). Adding syntactic information to LSA. In *Twenty-second Annual Conference of the Cognitive Science Society* (pp. 989-993). Mahwah, NJ: Lawrence Erlbaum Associates.

Abstract: Much effort has been expended in the field of Natural Language Understanding in developing methods for deriving the syntactic structure of a text. It is still unclear, however, to what extent syntactic information actually matters for the representation of meaning. LSA (Latent Semantic Analysis) allows you to derive information about the meaning without paying attention even to the order of words within a sentence. This is consistent with the view that syntax plays a subordinate role for semantic processing of text. But LSA does not perform as well as humans do in discriminating meanings. Can syntax be the missing link that will help LSA? This paper seeks to address that question.

Wild, F., Stahl, C., Stermsek, G., Penya, Y., & Neumann, G. (2005). Factors influencing effectiveness in automated essay scoring with LSA. In *Proc. of the 12th International Conference on Artificial Intelligence in Education (AIED)* (pp. 485-494). IOS Press.

Wild, F., Stahl, C., Stermsek, G., & Neumann, G. (2005). Parameters driving effectiveness of automated essay scoring with LSA. In *Proc. of the 9th International Computer Assisted Assessment Conference (CAA)* (pp. 485-494).

Wolfe, M. B. W., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W. et al. (1998). Learning from text: Matching readers and texts by latent semantic analysis. *Discourse Processes, 25,* 309-336.

Abstract: Examined the hypothesis that the ability of a reader to learn from text depends on the match between the background knowledge of the reader and the difficulty of the text information. Latent Semantic Analysis (LSA), a statistical technique that represents the content of a document as a vector in high-dimensional semantic space based on a large text corpus, was used to predict how much 106 college and medical student readers would learn from texts based on the estimated conceptual match between their topic knowledge and the text information. Ss completed tests to assess their knowledge of the human heart and circulatory system, then read 1 of 4 texts that ranged in difficulty from elementary to medical school level, and finally, completed the tests again. Results show a nonmonotonic relation in which teaming was greatest for texts that were neither too easy nor too difficult. LSA proved as effective at predicting teaming from these texts as traditional knowledge assessment measures. For these texts, optimal

assignment of text on the basis of either prereading measure would have increased the amount learned significantly. (PsycINFO Database Record (c) 2002 APA, all rights reserved)

Wolfe, M. B. W. & Goldman, S. R. (2003). Use of latent semantic analysis for predicting psychological phenomena: Two issues and proposed solutions. *Behavior Research Methods, Instruments & Computers, 35,* 22-31.

Abstract: Latent semantic analysis (LSA) is a computational model of human knowledge representation that approximates semantic relatedness judgments. Two issues are discussed that researchers must attend to when evaluating the utility of LSA for predicting psychological phenomena. First, the role of semantic relatedness in the psychological process of interest must be understood. LSA indices of similarity should then be derived from this theoretical understanding. Second, the knowledge base (semantic space) from which similarity indices are generated must contain "knowledge" that is appropriate to the task at hand. Proposed solutions are illustrated with data from an experiment in which LSA-based indices were generated from theoretical analysis of the processes involved in understanding two conflicting accounts of a historical event. These indices predict the complexity of subsequent student reasoning about the event, as well as hand-coded predictions generated from think-aloud protocols collected when students were reading the accounts of the event. (PsycINFO Database Record (c) 2003 APA, all rights reserved)

Yu, C., Cuadrado, J., Ceglowski, M., & Payne, J. S. (4-2-2005). Patterns in Unstructured Data - Discovery, Aggregation, and Visualization.

Ref Type: Unpublished Work

Yukawa, T., Kasahara, K., Kato, T., & Kita, T. (2001). An Expert Recommendation System Using Concept-Based Relevance Discernment. In *International Conference on Tools with Artificial Intelligence*.

Zampa, V. & Lemaire, B. (2002). Latent Semantic Analysis for User Modelling. *Journal of Intelligent Information Systems, 18,* 15-30.

Zha, H. & Simon, H. D. (1999). On Updating Problems in Latent Semantic Indexing. *SIAM Journal on Scientific Computing, 21,* 782-791.

No Author:

Overview of Interactive Learning Environments Vol * No 2 August 2000 (2000). [On-line]. Available:

http://www.swets.nl/sps/journals/ile0802.html

Notes: Abstracts of all articles in this special issue on LSA

Introducing the Intelligent Essay Assessor, web site does not exist anymore (07/02/06) (2001). [On-line].

Available: http://www.knowledge-technologies.com/products.html

Notes: Introduction to the IEA (webpage)

Exploratory Factor Analysis. (2005).

Ref Type: Unpublished Work

Latent Semantic Analysis (2005). [On-line]. Available: http://iv.slis.indiana.edu/sw/lsa.html

LSA Group Papers (2006). [On-line]. Available: http://lsa.colorado.edu/papers.html

Reference List. (2006).

Ref Type: Generic