How semantic is Latent Semantic Analysis?

Tonio Wandmacher

Laboratoire d'Informatique – E.A. 2101 – Equipe BdTln
Université François Rabelais de Tours
Site de Blois – 3 place Jean Jaurès – F-41000 Blois
tonio.wandmacher@etu.univ-tours.fr

Mots-clefs - Keywords

Analyse de la sémantique latente, analyse de collocations, relations lexicales, sémantique computationelle.

Latent Semantic Analysis, Collocation Analysis, lexical relations, computational semantics.

Résumé - Abstract

Au cours des dix dernières années, l'analyse de la sémantique latente (LSA) a été utilisée dans de nombreuses approches TAL avec parfois de remarquables succès. Cependant, ses capacités à exprimer des ressemblances sémantiques n'ont pas été réellement recherchées de façon systématique. C'est l'objectif de ce travail, où la LSA est appliquée à un corpus de textes de langue courante (journal allemand). Les relations lexicales entre un mot et ses termes les plus proches sont analysés pour un test de vocabulaire. Ces résultats sont alors comparés avec les résultats obtenus lors d'une analyse des collocations.

In the past decade, Latent Semantic Analysis (LSA) was used in many NLP approaches with sometimes remarkable success. However, its abilities to express semantic relatedness were not yet systematically investigated. This is the aim of our work, where LSA is applied to a general text corpus (German newspaper), and for a test vocabulary, the lexical relations between a test word and its closest neighbours are analysed. These results are compared to the results from a collocation analysis.

1 General introduction

In its beginnings, Latent Semantic Analysis aimed at improving the vector space model in information retrieval. Its abilities to enhance retrieval performance were remarkable; results could be improved by up to 30%, compared to a standard vector space technique (Dumais, 1995). It was further found that LSA was able to retrieve documents that did not even share a single word with the query but were rather semantically related.

This finding was the headstone for many subsequent researches. It was tried to apply the LSA approach to other areas, such as automated evaluation of student essays (Landauer et al., 1997) or automated summarization (Wade-Stein & Kintsch, 2003). In (Landauer & Dumais, 1997), even an LSA-based theory of knowledge acquisition was presented.

In these works, many claims on the analytic power of LSA were made. It is asserted that LSA does not return superficial events such as co-occurrence relations, but is able to describe semantic similarity between two words.[1] The extracted word relations are referred to as latent, hidden or deep[2], however, none of these papers addresses the nature of this deepness. LSA is called "semantic", but a thorough evaluation of its abilities to extract the semantics of a word or a phrase is missing.[3] One work that takes a little step in this direction, was done by Landauer & Dumais (1997). They use LSA-based similarities to solve a synonym test taken from the *TOEFL* (Test Of English as a Foreign Language) They found that the abilities of LSA to assign the right synonym (out of 4 test words) to the target word are comparable to those of human non-native speakers of English (mean LSA: 64,4%; mean humans: 64,5%).

However, this result can only be seen as a first indication for the capacity of LSA; it is neither a systematic assessment, nor a comparison to similar techniques. This is what we try to achieve in the following. Our aim is therefore not improvement, but evaluation and a better understanding of the method.

2 Presentation of LSA

LSA, as presented by (Deerwester et al. 1990) and others, is based on the vector space model of information retrieval (Salton & McGill, 1983). First, a given corpus of text is transformed into a term×context-matrix, displaying the occurrences of each word in each context. A context can be only a 2-word window, a sentence, a paragraph or a full text. For LSA, a paragraph window is normally assumed (cf. (Dumais, 1995), (Landauer et al, 1997)).

In a second step, this matrix is weighted by one of the weighting methods used in IR (c.f. (Salton & McGill, 1983)). For LSA, a log-entropy scheme showed the best results (Dumais, 1990). The decisive step in the LSA process is then a *Singular Value Decomposition* (*SVD*) of the weighted matrix. Thereby the original matrix A is decomposed as follows:

$$\mathrm{SVD}(A) = U \Sigma V^{\mathrm{T}} \qquad (1)$$

The matrices U and V consist of the eigenvectors of the columns and rows of A. Σ is a diagonal matrix containing in descending order the singular values of A. By only keeping the k

[1] Cf. (Wade-Stein & Kintsch, 2003), p. 10: "*LSA does not reflect the co-occurrence among words but rather their semantic relatedness.*"

[2] Cf. (Landauer et al., 1998), p. 4: "*It is important to note from the start that the similarity estimates derived by LSA are not simple contiguity frequencies, co-occurrence counts, or correlations in usage, but depend on a powerful mathematical analysis that is capable of correctly inferring much deeper relations.*"

[3] The 'latency' of LSA was indeed assessed by Wiemer-Hastings (1999).

strongest (*k* usually being around 300) singular values and remultiplying Σ_k with either *U* or *V*, one can construct a so-called *semantic space* for the terms or the contexts, respectively. Each term or each context then corresponds to a vector of *k* dimensions, whose distance to others can be compared by a standard vector distance measure. In most LSA approaches the cosine measure is used. By calculating the cosine of the angle between one term vector and all the others, a ranked list of next neighbours can be obtained for a given word. From the LSA point of view, these neighbours should be semantically related to the test word.

3 Method

To assess the abilities of LSA to generate semantic similarity, we applied it to a large corpus of German newspaper text. We used a random sample of 120.000 paragraphs (app. 20 mio. words) of the *Tageszeitung* (*TAZ*) from 1989 to 1998, which was stoplisted for frequent words and lemmatized by the *DMOR* package (Schiller, 1995). Words having a corpus frequency of less than 5 were also removed. This reduced the vocabulary size from 385.344 to 63.651 types. This was necessary, since the calculation of the SVD is heavily constrained by complexity matters.

After transforming the corpus into a term×context matrix (having the size 63.651 × 120.000), we applied a log-entropy weighting scheme. Using Michael Berry's GTP package v. 3.0 for Linux, we calculated the SVD for the above matrix up to 400 dimensions. To find the optimal factor *k*, we conducted some preliminary tests with various dimensionalities (250 – 400). A *k* of 309 gave the best results (in terms of the percentage of meaningful relations), even though the results for each of the samples were very close.

For a random sample of 400 words (nouns, verbs and adjectives only), their 20 next neighbours (= words having the highest cosine score with the centroid, see 2.) were extracted. We considered a fixed number of neighbours, since the usage of a threshold distance (e.g. *cos* = 0,5) proved not to be practical (the cluster size varied strongly).

The relations between the centroid (test word) and each of the 20 neighbours were then manually categorized in one of eight relation classes. The classes were the following:

- Synonymy
- Antonymy
- Hypo-/Hypernymy
- Co-Hyponymy
- Mero-/Holonymy

} truly semantic relations

- Loose association
- Morphological relation
- Erroneous relation

The notion of semanticity described by this classification can be questioned. However, our selection of semantic relations seems to be widely accepted in lexical semantics (cf. (Cruse, 1986)) and precise definitions exist to determine if a relation holds between two words *X* and *Y* (e.g. for meronymy: *X* IS-PART-OF *Y*). The same is true for the class of morphological

relations. A derivational or inflectional relation between two terms can be recognized easily most of the time.

Still, we admit that our classification is neither exhaustive, nor always clear-cut. Especially the class of "loose association" is rather intuitive. It was assumed as a collection class for all term pairs that were not related by definition of a semantic or a morphological relation, but still were somehow connected. Typical examples for this class might be '*Flugzeug*' ('airplane') / '*landen*' ('to land') or '*Katze*' ('cat') / '*Milch*' ('milk').

To balance out doubtful cases, we set the size of our test sample sufficiently large (20 neighbours for 400 words = 8000 categorized relations[4]) and had the classification task done independently by two German native speakers (including the author).

For each of the neighbours, additional information, such as its corpus frequency, context frequency and entropy value, was also determined.

4 Results

4.1 Quantitative Analysis

Results were calculated for the first 5, 10, 15 and 20 neighbours, respectively. As the fractions for each of the semantic classes were all quite low (0-5%), only the total of semantic relations is displayed here:

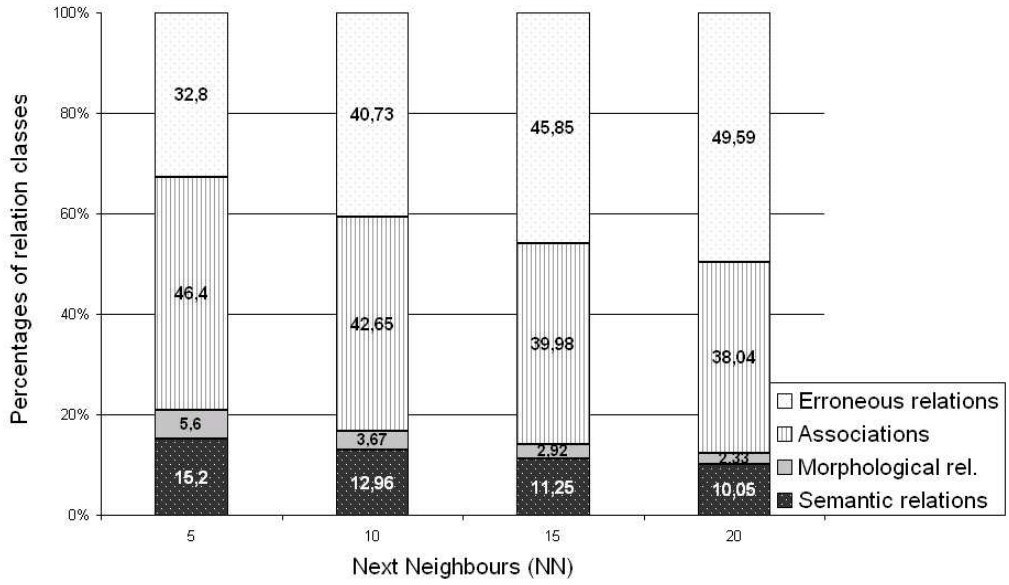

Figure 1: Percentages of relation classes resulting from LSA for a sample of 400 test words.

[4] For a sample size n = 400 the 95% confidence interval is maximally ± 4,9.

As the number of neighbours under consideration rises, the results get generally worse. Considering only the first five neighbours for every test word, we find nearly a third of erroneous relations (32,8%). Almost half of them are loose associations, whereas the truly semantic relations make up only 15%. Moreover, 5% of morphologically related word pairs were found.

When we consider the relations for 20 neighbours, the part of erroneous relations rises to nearly 50%, whereas the class of associations falls to 38%. Only 10% of the relations can be classified as semantic and app. 2% as morphological.

Splitting up the sample into parts of speech, we get the following picture:

Relation class	**Nouns**	**Adjectives**	**Verbs**
Semantic relations:	15,1%	7,8%	3,8%
Morphological relations:	2,4%	1,8%	2,1%
Associations:	39,3%	30,5%	41,0%
Erroneous relations:	43,2%	59,9%	53,1%

Table 1: Percentages of relations for the three parts of speech.

Table 1 shows a clear distinction. Nouns have far more meaningful relations (56,8%) than adjectives (40,1%) or verbs (46,9%). The difference becomes even clearer if only the class of semantic relations is considered. Opposing verbs and adjectives, another remarkable difference can be found: Verbs have much more (10,5%) associations than adjectives, but only less than half of semantic relations.

4.2 Qualitative Analysis

If one opposes the words with the lowest and the highest fractions of meaningful relations, a difference in usage of the two groups can be observed:

Lowest fractions (0-5%)		**Highest fractions (95-100%)**	
Ansehen (‚image')	natürlich (‚natural')	Mediziner (‚health prof.')	singen (‚to sing')
Aufbruch (‚breakup')	teilen (‚divide')	Reporter (‚reporter')	sterben (‚to die')
Beispiel (‚example')	zahlreich (‚numerous')	Therapie (‚therapy')	studieren (‚to study')
Kasten (‚box')	zumuten (‚to expect of')	Wohnraum (‚living space')	Luftwaffe (‚air force')
Umstand (‚circumstance')	überstehen (‚to overcome')	Zuhörer (‚auditor')	Malerei (‚painting')
Unsinn (‚nonsense')	Auflösung (‚resolution')	deportieren (‚to deport')	Religion (‚religion')
aufrecht (‚upright')	Rücksicht (‚consideration')	gesund (‚healthy')	Uniform (‚uniform')
automatisch (‚automatic)	bescheren (‚to bring')	kochen (‚to cook')	Wirtschaft (‚economy')
glatt (‚flat'/‚smooth')	denken (‚to think')	lernen (‚to learn')	lesen (‚to read')
intensiv (‚intensive')	einfallen (‚to occur')	operieren (‚to do a surgery')	orthodox (‚orthodox')

Table 2: Words of the sample having the lowest and highest fractions of meaningful relations with the test word.

Regarding the two groups shown in table 2, it appears that the words with the worst results can occur in every context. Words like '*Beispiel*' ('example'), '*Unsinn*' ('nonsense') or '*denken*' ('to think') are not connected to a certain theme or a typical context. On the other

hand, the words having the highest scores are rather specific. This group comprises terms such as '*Mediziner*' ('health professional'), '*Malerei*' ('painting') or '*kochen*' ('to cook'). These terms have a typical context; they are bound to a particular topic.

To get a clearer picture of this kind of specificity, it seems reasonable to further analyse the distribution of the words in the corpus. From the research on information retrieval it is known that specific terms are better predictors and get therefore a higher weight (Spärck-Jones, 1972), (Salton & McGill, 1983). The relevant values for the weighting schemes in IR are normally the *term frequency* (*tf*), the *corpus frequency* (*cf*) and the *context* (or *document*) *frequency* (*df*). A combination of these values forms the base for nearly all weighting schemes of the so-called *tf*idf*-family (s. (Salton & McGill, 1983)). Could these values be good predictors for our purposes?

We calculated the correlation between several of these values (as well as some combinations) and the fraction of meaningful relations among 20 next neighbours, with interesting results: Neither the simple frequency measures (corpus and document frequency) nor the entropy of a term showed a significant correlation with the fraction of meaningful relations. While trying out several combinations of the values, we found only one that showed a slight correlation (*Pearson*-Coefficient = 0,32, significance level <0,001), namely the quotient of *cf* and *df*.

In addition, we calculated the correlation between the mean distance of the 20 neighbours and the percentage of meaningful relations for the whole test set. We observed a medium correlation (*Pearson*-Coeff. = 0,56 at a significance level of <0,001). We therefore can conclude that medium distance and relation quality are related, although not too strongly.

5 Comparison with collocation analysis

5.1 Method

To obtain a contrastive example, we did the same experiment using collocation analysis (CA). We hereby used a formula presented by Quasthoff & Wolff (1998), (2002). They calculate the collocative significance between two words *A* and *B* as follows:

$$sig(A,B)=\frac{C_A C_B}{n}-k\cdot\log\frac{C_A C_B}{n}+\log k! \qquad (2)$$

where C_A (C_B) is a context, in which *A* (*B*) occurs, *n* is the amount of all contexts and *k* the number of all contexts containing *A* and *B*.[5]

To ensure comparability, we used the same corpus and did the same pre-processing (i.e. stoplisting, removal of low frequency words etc.). We then calculated for our sample of 400 test words the 20 words having the highest collocative significance with the test word. This gave us again 400 word clusters, for which every relation was categorized as above.

[5] This measure is obviously related to the one given by Dunning (1993). Both measures appear to give rather similar results (cf. (Quasthoff & Wolff, 2002)).

5.2 Results

Figure 2 shows the fractions of the different relation classes obtained by collocation analysis:

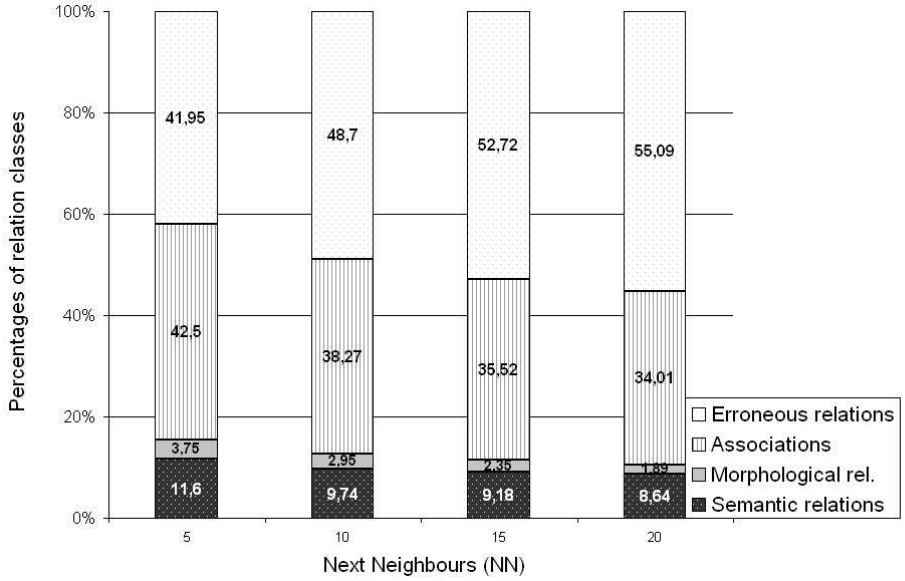

Figure 2: Percentages of relation classes resulting from CA for a sample of 400 test words.

Comparing figures 1 and 2, the results seem quite close. The differences between the fractions for the meaningful classes do not exceed 4,5%. This is remarkable, keeping in mind that the collocation analysis exploits only co-occurrence relations in the text.

As a whole, the percentage of meaningful relations is a little lower for CA than for LSA. Still we find 8% (20 NN) to 12% (5 NN) of semantic relations, not much less regarding the LSA results (10%-15%). Still, all LSA scores were significantly higher than the CA scores in a Student's T test (significance level: <0,001).

Regarding the words and their neighbours, we get a similar picture as with LSA: terms being bound to a particular theme get better results than those that are context-independent. Moreover, the results of CA and LSA show a high correlation (*Pearson*-Coeff. = 0,72 at a significance level of <0,001) for our sample. It seems therefore that LSA and CA make use of the same properties of a word and its context.

However, one difference can be observed with regard to ambiguous words. CA returns neighbours that still belong to both meanings, LSA however seems to mask out one of them. This behaviour can be seen in the following examples for '*Bach*' (Meanings: J. S. Bach, the composer / 'creek') and '*Schlange*' (Meanings: 'queue' / 'snake'):

	Next neighbours for '*Bach*' ('creek'/J.S. Bach)		Next neighbours for '*Schlange*' ('queue'/'snake')	
Rank	**LSA results**	**CA results**	**LSA results**	**CA results**
1.	Musik (‚music')	Sebastian (‚Sebastian')	stehen (‚to stand')	stehen (‚to stand')
2.	Beethoven (‚Beethoven')	Johann (‚John')	warten (‚to wait')	warten (‚to wait')
3.	musizieren (‚to make music')	Musik (‚music')	Schild (‚sign')	Reptil (‚reptile')
4.	klanglich (‚sonorical')	Emanuel (‚Emanuel')	Wartezeit (‚waiting time')	lang (‚long')
5.	musikalisch (‚musical')	Elvira (‚Elvira')	Kaufhaus (‚department store')	bilden (‚to form')
6.	Klang (‚sound')	Mozart (‚Mozart')	Supermarkt (‚supermarket')	Schalter (‚counter')
7.	Gesang (‚chant')	Artist (‚artist')	Vesna (‚name')	Kaninchen (‚rabbit')
8.	rhythmisch (‚rhythmical')	runtergehen (‚to flow down')	lang (‚long')	Buchstabe (‚letter')
9.	komponieren (‚to compose')	verunreinigen (‚to pollute')	Straßenrand (‚roadside')	Mensch (‚human')
10.	Improvisation (‚improvisation')	Brahms (‚Brahms')	Bäckerei (‚bakery')	giftig (‚poisonous')
11.	Mozart (‚Mozart')	Fluss (‚river')	tagsüber (‚in the day')	auftauchen (‚to emerge')
12.	virtuos (‚'virtuoso)	Ton (‚sound/note')	Matte (‚mat')	einreihen (‚to queue')
13.	Komposition (‚composition')	hinunter (‚down')	Stau (‚holdup')	Warteschlange (‚queue')
14.	Rhythmus (‚rhythm')	Gewässer (‚water')	Warteschlange (‚queue')	lange (‚long')
15.	Saxophon (‚saxophone')	Geige (‚violin')	Einlass (‚entry')	Australien (‚Australia')
16.	Geige (‚violin')	Oboe (‚oboe')	Brot (‚bread')	Stau (‚holdup')
17.	Komponist (‚composer')	Flussufer (‚river bank')	Greenfield (‚Greenfield')	Tag (‚day')
18.	akustisch (‚acustical')	Aufführung (‚performance')	lange (‚long')	Käfig (‚cage')
19.	klassisch (‚classical')	rauschen (‚rush')	Auslage (‚display')	öffnen (‚to open')
20.	Cello (‚cello')	Philipp (‚Philipp')	Mittelpunkt (‚center')	Wartezeit (‚waiting time')

Table 3: Two examples of ambiguous words and their next neighbours. The neighbours belonging to the prominent meaning are in blank, the ones of the non-prominent meaning are in black. Terms that cannot be assigned are shaded in grey.

The difference in the analyses is obvious: in both examples, LSA generates neighbours of the prominent meaning only, whereas the NN produced by CA are of both domains. However, this is only a first observation; we did not yet assess this masking-out property of LSA in a systematic way. It should be subject to further research.

6 Discussion

To take up our initial question: how semantic is LSA? The conclusion that we draw from our results, is: not as much as its name might suggest. The fractions of truly semantic relations were not very high (10% at the 20-NN level), and a big part (38% at 20 NN) of the relations, however, could rather be described as associative. These words are conceptually related, but not necessarily in a narrow semantic or morphological sense.

The biggest part however, nearly half of the relations generated at the 20-NN level, are erroneous, i.e. there is no apparent relation between the test word and its neighbour.

Comparing the LSA approach to other procedures exploiting co-occurrence information, one reason for the high percentage of error relations may be found: Techniques such as *HAL* (Lund & Burgess, 1996) or the approach by Rapp (2002), (2003) use a co-occurrence window of a few (3-40) words only. LSA however relies on full paragraphs (average length in our case: 102 words). And even though a paragraph can be regarded as a semantically coherent unit, many of the inter-word relations in it may already be too weak. This may cause arbitrary relations.

Another questionable point about LSA arises from the modelling itself. The term-by-context-matrix is extremely sparse. In our experiments, the matrix had only maximally 0,08% nonzero elements. This is by itself of course not harmful, but recalling that the complexity of the SVD process constrains the overall size of the matrix, a different modelling seems more reasonable. Again, the approaches of Lund and Burgess (1996) and Rapp (2003) may give the answer: A term-by-term-matrix[6] can model the same amount of text in a smaller and less sparse format. Using this kind of matrix, much larger corpora can be used for the analysis; Rapp (2003) was able to analyse the full *British National Corpus* comprising more than 100 million words.

With respect to the results obtained from a collocation analysis of the same corpus, the LSA results do not show big differences. In general, they are significantly better, but none of the classes differs more than 4,5% from the CA. This is surprising, since a technique like CA relies on co-occurrence information only and does not make use of complex matrix calculations.

Still, the two analyses seem to show a different behaviour with regard to ambiguous words. Whereas CA finds for a given ambiguous test word neighbours from both conceptual domains, LSA seems to mask out one of the meanings.

We hope to have given a deeper understanding of what LSA can and cannot do. Regarding our results, it is not much more semantic than a simple technique like CA, and some of its modelling aspects, such as the optimal context size or the kind of co-occurrence matrix still leave space for improvement and further research. Particularly the effects of the *SVD* on word similarity should be further investigated, before LSA is used as a general tool to derive semantic relations from text.

References

CRUSE, D.A. (1986), *Lexical Semantics*, Cambridge University Press, Cambridge.

DEERWESTER, S. C., DUMAIS, S.T., LANDAUER, T.K., FURNAS, G.W., HARSHMAN, R.A. (1990), "Indexing by Latent Semantic Analysis", *Journal of the American Society of Information Science*, 41 (6), pp. 391 – 407.

DUMAIS, S. (1990), *Enhancing Performance in Latent Semantic Indexing*, Technical Report TM-ARH-017527, Bellcore.

[6] A term-by-term matrix for a vocabulary V is of the size $|V|^2$ and reflects the frequency of co-occurrence of two terms within a certain text window (e.g. ± 5 words).

DUMAIS, S. T. (1995), "Latent Semantic Indexing (LSI): TREC-3 Report", in D. Harman (Ed.), *Proceedings of the 3rd Text REtrieval Conference (TREC-3)*, Vol. 500-226, pp. 219-230, NIST Special Publication.

DUNNING, T. (1993), "Accurate Methods for the Statistics of Surprise and Coincidence", *Computational Linguistics* **19(1)**, pp. 61-74.

LANDAUER, T. K. und DUMAIS, S. T. (1997), "A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge", *Psychological Review* **104**, pp. 211-240.

LANDAUER, T. K., LAHAM, D., REHDER, B., und 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 und P. Langley (Eds.), *Proceedings of the 19th annual meeting of the Cognitive Science Society*, pp. 412-417, Erlbaum, Mawhwah, NJ.

LANDAUER, T. K., FOLTZ, P. W., und LAHAM, D. (1998), "Introduction to Latent Semantic Analysis", *Discourse Processes* **25**, pp. 259-284.

LUND, K. and BURGESS, C. (1996), "Producing high-dimensional semantic spaces from lexical co-occurrence", *Behaviour Research Methods, Instruments and Computers* **28**(2), pp. 159-165.

QUASTHOFF, U. (1998), „Deutscher Wortschatz im Internet", in *Proceedings des LDV-Forum 2/98*.

QUASTHOFF, U. und WOLFF, C. (2002), "The Poisson Collocation Measure and its Applications", in *Proceedings of the 2nd International Workshop on Computational Approaches to Collocations*, Wien.

RAPP, R. (2002), "The Computation of Word Associations: Comparing Syntagmatic and Paradigmatic Approaches", *Proceedings of the COLING 02*, Taipei.

RAPP, R. (2003), "Word Sense Discovery Based on Sense Descriptor Dissimilarity", *Proceedings of the 9th Machine Translation Summit*, New Orleans.

SALTON, G. und MCGILL, M. (1983), *Introduction to Modern Information Retrieval*, McGraw-Hill, New York.

SCHILLER, A. (1995), *DMOR: Benutzeranleitung*, Technical report, IMS Stuttgart, Draft.

SPÄRCK-JONES, K. (1972), "A statistical interpretation of term specificity and its application in retrieval", *Journal of Documentation* **28(1)**, pp. 11-20.

WADE-STEIN, D. und KINTSCH, E. (2003), *Summary Street: Interactive Computer Support for Writing*, Technical report, University of Colorado.

WIEMER-HASTINGS, P. (1999) "How latent is Latent Semantic Analysis?", *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, Aug 1999, pp. 932-937. San Francisco: Morgan Kaufmann.