PARAMETERS DRIVING EFFECTIVENESS OF AUTOMATED ESSAY SCORING WITH LSA

Fridolin Wild, Christina Stahl, Gerald Stermsek
and Gustaf Neumann

Parameters Driving Effectiveness of Automated Essay Scoring with LSA

Fridolin Wild, Christina Stahl, Gerald Stermsek, Gustaf Neumann
Department of Information Systems and New Media,
Vienna University of Economics and Business Administration,
Augasse 2-6, A-1090 Vienna, Austria
{firstname.lastname}@wu-wien.ac.at

Abstract

Automated essay scoring with latent semantic analysis (LSA) has recently been subject to increasing interest. Although previous authors have achieved grade ranges similar to those awarded by humans, it is still not clear which and how parameters improve or decrease the effectiveness of LSA. This paper presents an analysis of the effects of these parameters, such as text preprocessing, weighting, singular value dimensionality and type of similarity measure, and benchmarks this effectiveness by comparing machine-assigned with human-assigned scores in a real-world case. We show that each of the identified factors significantly influences the quality of automated essay scoring and that the factors are not independent of each other.

Introduction

Computer assisted assessment in education has a long tradition (cf. Hearst, 2000). Early experiments on grading free text responses were conducted during the Project Essay Grade (PEG) by Page (Page, 1966) in 1966. Page used punched cards and based his experiments predominantly on syntactical features. Research today focuses on emulating a human-semantic understanding, backed up by hitherto unknown computing power. Semantic understanding in automated essay scoring is required in order to automatically evaluate the content of written essays on the basis of predefined text corpora. According to Whittington and Hunt (Whittington and Hunt, 1999), however, free text assessment is a complex and fundamentally subjective process. Correspondingly, several findings report the correlation in grade attribution between two human assessors to be located around .6 (cf. e.g. Landauer & Psotka, 2000).

Landauer et al. (Landauer et al., 1998) combined the vector model of a document-term-space with singular value decomposition (SVD, a two mode factor analysis) – a method they named ‘latent semantic analysis’ (LSA). This approach reveals the latent semantic structure of text corpuses, while eliminating noise in word application. Using LSA to assess written essays enables grade ranges similar to those awarded by human graders. Several stages in this process leading from raw input documents to the machine assigned scores allow for optimisations, many of them in common with other application areas of latent semantic analysis and indexing.

The process of auto-scoring can be divided into five sub-steps: *text pre-processing*, *weighting*, *calculation of the SVD*, *correlation measurement* and *correlation method*. Contradicting claims can be found concerning the adjustment of influencing factors in these steps.

Perfetti (Perfetti, 1998), for example, argues for more reliable results with a larger corpus size (input documents). On the other hand, Deerwester et al. (Deerwester et al., 1990) were able to successfully apply LSA to a corpus with only nine documents. Nakov (Nakov, 2000) reports the best results with a raw term frequency applied as local weighting scheme, whereas Dumais (Dumais, 1990) finds a logarithmized term frequency suiting best. Dimensionality is seen quite different by authors. Dumais (Dumais, 1990) sticks to the magical number of 100, whereas Graesser et al. (Graesser et al., 1999) suggest the use of 100 to 300 dimensions. Nakov (Nakov, 2000) recommends the number of dimensions to vary from 50 to 1500. Our own experiments have shown that the 10 dimensions are often a good starting point.

Conclusions on how to calibrate an LSA essay scoring process can hardly be drawn from these statements. Furthermore, these examples indicate that identifying the perfect calibration is complex and tightly coupled to the purpose the application serves.

In this contribution we describe and conduct an experiment on varying influence factors and their optimisation within the application of essay scoring using LSA. While in our first paper on this issue (to be published) we varied all influence factors ceteris paribus, in this paper we analyse the mutual influence of factors by investigating all 2016 possible combinations. As a side effect, we provide a proof-of-concept for a real world case with relatively small text-corpora in German. It is not our goal to automatically calculate grades, but to investigate the parameters driving the automated scoring of free text answers with LSA as a basis for automatic feedback and artificial tutoring.

The rest of the paper is structured as follows. In Section 1 we explicate what we understand by automated essay scoring and expose the algorithm and the setting with which the LSA and the scoring process were performed. The methodology applied in this research is documented in Section 2. In Section 3 we explain our hypothesis and describe the test design of our experiments. The resulting data is analyzed in Section 4, where we also review our hypothesis. Section 5 gives an outlook on future research.

Automated Essay Grading with Latent Semantic Analysis (LSA)

In opposite to multiple-choice tests, this paper addresses means of how to conduct (auto-)assessment of free text responses with latent semantic analysis (LSA). Hereby the problem of ‘feeding’ students the right answers, rather than actually testing their active knowledge, can be avoided. In accordance with Stalnaker (Stalnaker, 1951) we define an essay to be “…a test item which requires a response composed by the examinee, usually in the form of one or more sentences, of a nature that no single response or pattern of responses can be listed as correct”.

Automatic essay scoring furthermore means, that not only a human person is capable of performing the marking, but – beyond Stalnaker – that a machine can successfully emulate human scoring. In our view this requires an expert skilled and informed in the subject to invest knowledge, and a linguistic engineer to set up a system that is capable of using this represented knowledge to assess incoming essays in respect to their quality and accuracy.

Derived from latent semantic indexing, LSA is intended to enable the analysis of semantic structure of texts (Deerwester et al., 1990). The basic idea behind LSA is that a collocation of terms of a given document-term-space reflects a higher-order – latent semantic – structure, which is obscured by word usage (e.g. synonyms or ambiguities). By using conceptual indices that are derived statistically via truncated singular value decomposition, this variability problem can be overcome (cf. Berry et al., 1995).

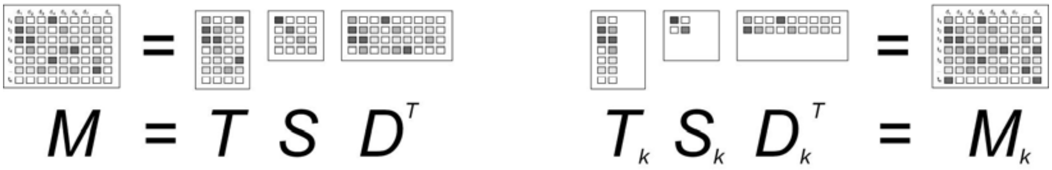

Figure 1. Singular Value Decomposition (original left, truncated right)

A document-term-matrix is constructed from a given text base of *n* documents containing *m* terms (see *M* in Figure 1). This matrix *M* of the size m×n is then decomposed via singular value decomposition into: term vector matrix *T* (constituting left singular vectors), the document vector matrix *D* (constituting right singular vectors) being both orthonormal and the diagonal matrix *S* (constituting singular values). Multiplying the truncated matrices T_k, S_k and D_k results in a new matrix M_k which is the least-squares best fit approximation of *M* with *k* singular values.

In case column 1 of M_k constitutes the document vector of a perfect essay (a ‘golden’ essay) and column 6 of M_k were an essay to be graded, the grade could be calculated as the correlation between these two vectors.

To keep the essays that are to be tested from influencing the factor distribution of the underlying text corpus, they are folded-in after the svd decomposition. Folding-in is done by calculating a pseudo document vector m_i in M_k of the essay to be tested as shown in equation 1 and 2 (cf. Berry et al., 1995):

1) $d_i = v^T T_k S_k^{-1}$

2) $m_i = T_k S_k d_i^T$

The document vector v^T in equation 1 is identical to an additional column of *M* with the term frequencies of the essay to be folded-in. T_k and S_k are the truncated matrices from an SVD applied on a given text collection used to construct the latent semantic space. For essay scoring the text collection usually consists of text book sections, model answers, glossary entries, generic texts, and the like (cf. Deerwester et al., 1990; Graesser et al., 1999).

Methodology

A software programme written in the statistical programming language R was used for automatic essay scoring, enabling us to alter the presumed influencing factors we selected in our experimental design. An experiment investigates cause-and-effect relationships where independent variables are modified to analyze their effects on dependent variables (Picciano, 2004), which we did by changing the influencing factors (our independent variables). Our dependent variables were the machine assigned essay scores which we compared to the human assessed scores by measuring their correlation.

For our experiments we used a German text corpus of students' answers to a marketing question taken from a previous real world exam. The corpus was composed of 43 files, which had been pre-graded by a human assessor with points from 0 to 5. Points – other than grades – are ratio-scaled variables with equal intervals and an absolute zero point.

We took three 'golden essays' from the text corpus to compare them to the remaining essays. To build the semantic space we used a marketing glossary of 302 files, each file containing one glossary entry. The glossary is part of the marketing course material offered via our university's e-learning platform Learn@WU. On average the essays had 56.4 words and the glossary entries had a length of 56.1 words.

Hypothesis and Test Design

We performed numerous tests to investigate the impact of five aspects that have shown great influence on the effectiveness of LSA: (1) the preprocessing of the input text, (2) the use of weighting-schemes, (3), the choice of dimensionality, (4) the applied similarity measure and (5) the applied similarity measurement method (see Figure 2).

To understand how these aspects affect each other, we tested every possible combination of the four influencing factors resulting in 2016 test runs. Thus, we hope to find the best setting for every influencing factor based on the impact of all other influencing factors.

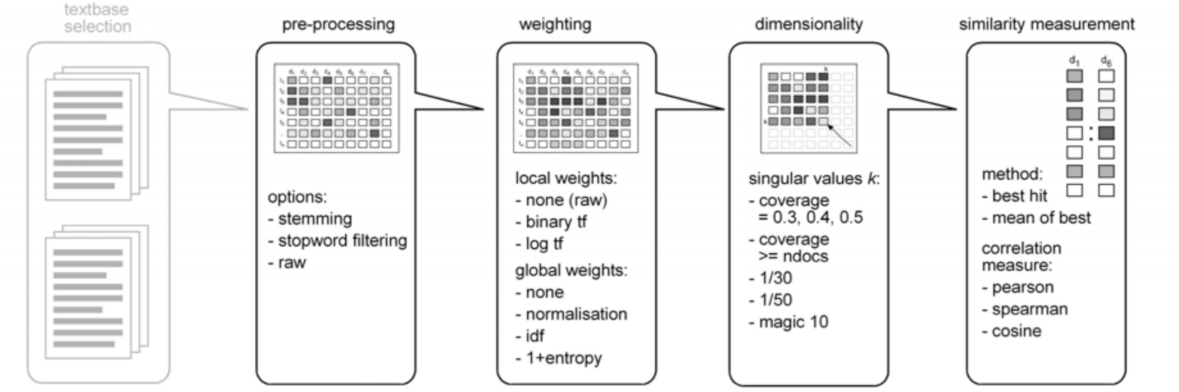

Figure 2. Considered influencing factors

Document pre-processing

Document pre-processing is a widely used procedure in information retrieval. Typical text operations include lexical analysis, use of stop-word lists, stemming, selection of index terms, and construction of thesauri (Baeza-Yates and Ribeiro-Neto, 1999).

Our tests focused on the elimination of stop-words and stemming. Because document pre-processing in general improves results for information retrieval its advantages for LSA seem straightforward. In their experiments Nakov et al. (Nakov et al., 2001) indeed obtained better overall results when removing stop-words but a significant improvement for stemming was only found in one case. This supports the hypotheses that the removal of stop-words improves LSA, whereas the use of stemming-algorithms has only little influence.

To assess the effects of pre-processing we used a stop-word list with 373 German terms and Porter’s Snowball stemmer to remove common word suffixes. We tested the corpus with stop-word removal and stemming, stop-word removal only, stemming only and no pre-processing at all resulting in four different settings for the document pre-processing step.

Weighting-Schemes

Weighting-schemes have the most extensive impact on the effectiveness of LSA. Several weighting-schemes – both local and global – have been tested for various languages and applications of LSA yielding best results for the logarithm as the local, and the entropy as the global weighting (e.g. Dumais, 1990; Nakov et al., 2003).

We have no knowledge, however, about the way weighting-schemes affect document pre-processing, the choice of dimensionality and the similarity measures. To address this matter we combined three local (raw term-frequency, logarithm, and binary) and three global (normalization, inverse document-frequency, and entropy) weightings. We also performed tests without any global weighting leading to 3 x 4 = 12 different weighting settings all together.

Choice of Dimensionality

Among other factors the choice of dimensionality has a significant impact on the results of LSA. After calculating the singular value decomposition from the original term-document matrix, a reduced matrix is reconstructed using only the k-largest singular values. This aims at obtaining an approximation to the original vector space, which captures the most important structure but reduces noise and variability in word usage (Berry et al., 1995). For our tests we considered the following alternatives to determine the number of selected factors in the approximated vector space:

- *Percentage of cumulated singular values (share)*: Using a normalized vector of cumulated singular values we can sum up singular values until we reach a specific value. In our paper we suggest to use 50%, 40% and 30% of the cumulative summed up singular values.
- *Absolute value of cumulated singular values equals number of documents (ndocs)*: The sum of the first k singular values factors equals the number of documents n used in the analysis.
- *Percentage of number of terms (1/30, 1/50)*: Alternatively the number of factors can be determined by a fraction of all terms indexed. Commonly used fractions are 1/30 or 1/50.
- *Fixed number of factors (magic 10)*: A less sophisticated and inflexible approach is to use a fixed number of factors, e.g. 10 factors. The number has to be determined depending on the text corpus.

Similarity Measures & Methods

Finally, we tested three similarity measures: Pearson-correlation, Spearman’s rho and Cosine. Although the cosine is the most commonly used measure for comparing LSA-vectors and usually works best for information retrieval (Landauer and Dumais, 1997) our past experiments returned the best result for Spearman’s rho. Additionally we investigated the different values for both, the correlation with the best matching golden essay (*max*) and the mean correlation of all three golden essays (*mean*) resulting in 2 x 3 = 6 different settings.

Reporting Results

The results of our test series can be seen in the following Figures 1–5. With an average Spearman correlation of 0.31, filtering stop-words without stemming turned out to produce overall better results than the other pre-processing methods. Interestingly, stop-word filtering compressed the correlation curves, so that it starts with a higher offset and leads to higher maxima (see Figure 1). Contrarily, stemming and the combination of stop-word filtering with stemming both reduced the average correlation with the human scores. On average stemming alone produced by .06 and stemming combined with stop-word filtering by .03 worse results than the raw document-term-matrices with an average correlation of .26. Accordingly within 50 best combinations, stopping could be found 21 times, raw matrices 14 times, stemming 12 times and stemming combined with stopping 3 times.

Within the term weighting algorithms, the inverse document frequency (IDF) proofs to be the best of the global weighting schemes, whereas the local weighting schemes on average seem to have hardly any effect. However, as can be seen in Figure 2, using the raw or the logarithmized local term frequencies compresses the corresponding correlation curves and even the significance levels more than the binary term frequency.

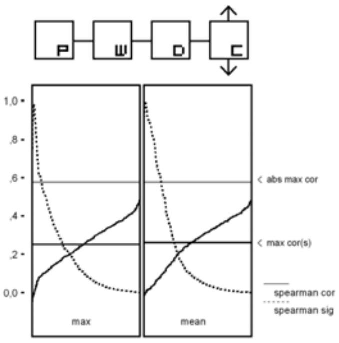

Figure 5. Test results sorted by correlation method

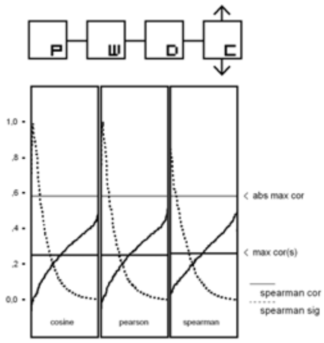

Figure 4. Test results sorted by correlation measure

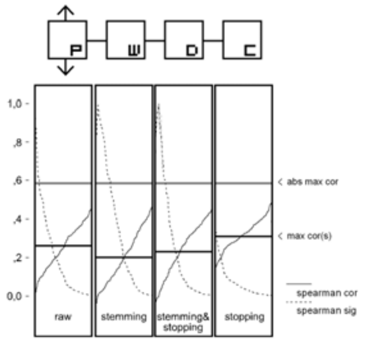

Figure 1. Test results sorted by pre-processing method

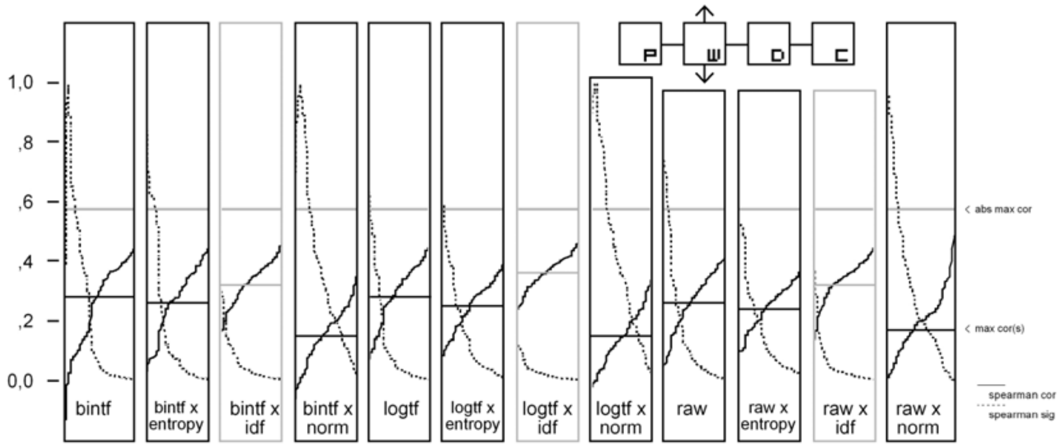

Figure 2. Test results ordered by applied weighting scheme

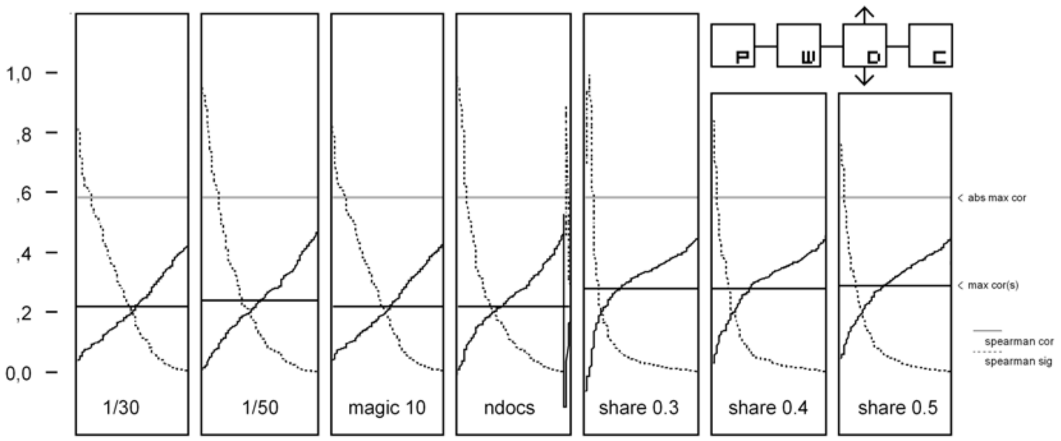

Figure 3. Correlations & p-values for 2016 tests, sorted by dimension calculus

After all, the logarithmized term frequencies combined with IDF yields with .36 the best average results. Normalisation has shown to produce overall worse results in all test runs (with .15 average for binary term frequencies, .15 average for logarithmized term frequencies and .17 average for raw term frequencies). Entropy (‘1 + entropy’ to be precise) seems to have nearly no effect at all, performing minimally lower on average than unprocessed term frequencies. The 50 best combinations of influencing factors returned 20 times binary term frequencies, 19 times logarithmized frequencies and 11 times the raw frequencies as the best local weighting scheme. The global weighting schemes were 26 times IDF, 13 times raw, 6 times normalisation and 5 times entropy.

Among the different methods to determine the number of singular values, the group percentage of cumulated singular values creates on average and overall better correlations with the human scores (see Figure 3). Of them, the share of 50% produces with .29 slightly better results than going for 40% or 30% (both .28). However, the distribution of correlation and significance values favors the use of the 30% rule. The other methods look very similar and vary from .22 to .24 in their average Spearman correlations1. The negative correlations on the left hand side of share 0.3 are combinations using normalisation (followed by entropy) for global weighting. Within the best combinations, 1/50th proved with 13 times to be good, followed by 10 times share 50%, 8 times 1/30th, 8 times magic 10, 5 times share 40%, 3 times share 30% and 3 times ndocs.

Looking at the results of the test runs sorted by the applied correlation measure as depicted in Figure 4, the Cosine and Pearson correlation perform slightly better than Spearman’s rho (although on average they both score only .25 compared to .26 of the Spearman correlation). However, as can be clearly seen from the curves, they reach approximately the same maximum and – despite they have a smaller minimum – their curves vault a little more in the higher correlation areas. However, within the best test runs, the Spearman correlation was found 21 times, followed by 15 times Cosine and 14 times Pearson. Comparing the two methods ‘best-match with one golden essay’ and the ‘average of the match with three golden essays’, the mean scored slightly better on average (with .26 compared to .25). Within the best combinations, mean even outperformed more with 31 times compared to only 19 times for the maximum best match method.

Besides Spearman’s rho, we also measured the Kendall correlation between the machine-assigned and the human-assigned scores which detected slightly lower correlations. The trend of each graph, however, was identical.

[1] The outlier to the right of the ndocs-diagram is result of a calculatory problem: in some cases the number of singular values not breaking the number of documents with their summed up values turns out to be only one. In this case it is not possible to calculate the matrix multiplications as given in the formulas (1) and (2). To visually separate them from the other values, they have been sorted to the right of the graphic.

Conclusions and Future Work

In this paper we investigated the influencing factors on effectiveness in automatic essay scoring. Our results give evidence, that for the real world case we tested, the identified parameters drive the correlation of the machine assigned with the human scores.

However, several recommendations on the adjustment of these parameters, which we extracted from literature, do not apply in our case. Through our large scale experiment with 2016 test runs, we found evidence that stop-word filtering is the best performing method within the pre-processing step. Within the weighting step, we detected the inverse document frequency as outperforming global weighting scheme and no clear preference for a local one. Spearman's rho could be identified to be the most promising correlation measure, and the mean match against all three golden essays was found to suite the measurement step in most cases best.

As most of the influencing factors potential settings were actually part of the best performing combinations, we conclude that optimisations are not independent of each other. Furthermore, we suspect that their adjustment strongly relies on the corpus used as text base and on the essays to be assessed.

Nevertheless, significant correlations between machine and human scores could be discovered, which ensures, that the applied LSA method can be exploited to automatically create valuable feedback on learning success and knowledge acquisition. We were able to work out recommendations on which parameter settings in general enhance performance.

Based on these results, we intend to investigate the stability of our results within the same discipline and in different contexts. Additionally, we want to elaborate a model indicating in detail the dependencies between different influencing factors, thus enabling us to successfully predict effectiveness. Furthermore, we intend to investigate scoring of essays not against best-practice texts, but against single aspects, as this would allow us to generate a more detailed feedback on the content of essays.

References

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

Berry, M., et al. (1995): Using Linear Algebra for Intelligent Information Retrieval. In: SIAM Review 37(4), pp. 573-595.

Deerwester, S., et al. (1990): Indexing by Latent Semantic Analysis. In: Journal of the American Society for Information Science 41(6), pp. 391-407.

Dumais, S. (1990): Enhancing Performance in Latent Semantic Indexing (LSI) Retrieval. Technical Report, Bellcore.

Graesser, A., et al. (1999): AutoTutor: A simulation of a human tutor. In: Journal of Cognitive Systems Research 1, pp. 35-51.

Hearst, M. (2000): The debate on automated essay grading, In: IEEE Intelligent Systems, 15(5), pp. 22-37

Landauer, T., Dumais, S. (1997): A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. In: Psychological Review 104 (2), pp. 211-240.

Landauer, T., et al. (1998): Introduction to Latent Semantic Analysis. In: Discourse Processes 25, pp. 259-284.

Landauer, T., Psotka, J. (2000): Simulating Text Understanding for Educational Applications with Latent Semantic Analysis: Introduction to LSA. In: Interactive Learning Environments 8 (2), pp. 73-86.

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., et al. (2001): Weight functions impact on LSA performance. In: Recent Advances in Natural language processing – RANLP'2001. Tzigov Chark, Bulgaria, pp. 187-193.

Nakov, P., et al. (2003): Towards Deeper Understanding of the LSA Performance. In: Recent Advances in Natural language processing – RANLP'2003, pp. 311-318.

Page, E. (1966). The imminence of grading essays by computer. Phi Delta Kappan 47, 238-243.

Perfetti, C. (1998): The Limits of Co-Occurrence. In: Discourse Processes 25(2&3), pp. 363-377.

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

Stalnaker, J. M. (1951): The Essay Type of Examination. In E. F. Lindquist (Ed.): Educational Measurement. (pp. 495-530). Menasha, George Banta.

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