# Positioning for Conceptual Development using Latent Semantic Analysis

**Fridolin Wild, Bernhard Hoisl**
Vienna University of Economics
and Business Administration

**Gaston Burek**
University of Tübingen
Computational Linguistics Division

## Abstract

With increasing opportunities to learn online, the problem of positioning learners in an educational network of content offers new possibilities for the utilisation of geometry-based natural language processing techniques.

In this article, the adoption of latent semantic analysis (LSA) for guiding learners in their conceptual development is investigated. We propose five new algorithmic derivations of LSA and test their validity for positioning in an experiment in order to draw back conclusions on the suitability of machine learning from previously accredited evidence. Special attention is thereby directed towards the role of distractors and the calculation of thresholds when using similarities as a proxy for assessing conceptual closeness.

Results indicate that learning improves positioning. Distractors are of low value and seem to be replaceable by generic noise to improve threshold calculation. Furthermore, new ways to flexibly calculate thresholds could be identified.

## 1 Introduction

The path to new content-rich competencies is paved by the acquisition of new and the reorganisation of already known concepts. Learners willing to take this journey, however, are imposed with the problem of positioning themselves to that point in a learning network of content, where they leave their known trails and step into the unknown – and to receive guidance in subsequent further conceptual development.

More precisely, positioning requires to map characteristics from a learner's individual epistemic history (including both achievements and shortcomings) to the characteristics of the available learning materials and to recommend remedial action on how to achieve selected conceptual development goals (Van Bruggen et al., 2006).

The conceptual starting points of learners necessary to guide the positioning process is reflected in the texts they are writing. Through structure and word choice, most notably the application of professional language, arrangement and meaning of these texts give cues about the level of competency[1] development.

As learning activities increasingly leave digital traces as evidence for prior learning, positioning support systems can be built that reduce this problem to developing efficient and effective matchmaking procedures.

Latent semantic analysis (LSA) (Deerwester et al., 1990) as one technology in the family of geometry-based natural language models could in principle provide a technological basis for the positioning aims outlined above. The assumption underlying this is that the similarity to and of learning materials can be used as a proxy for similarity in learning outcomes, i.e. the developmental change in conceptual coverage and organisation caused by learning.

In particular, LSA utilises threshold values for the involved semantic similarity judgements. Traditionally the threshold is obtained by calculating the average similarity between texts that correspond to the same category. This procedure can be inaccurate if a representative set of documents for each category is not available. Furthermore, similarity values tend to decrease with increasing corpora and vocabulary sizes. Also, the role of distractors in this context, i.e. negative evidence as reference material to sharpen classification for positioning, is largely unknown.

With the following experiment, we intend to

[1] See (Smith, 1996) for a clarification of the difference of competence and competency

*Proceedings of the EACL 2009 Workshop on GEMS: GEometical Models of Natural Language Semantics*, pages 41–48,
Athens, Greece, 31 March 2009. ©2009 Association for Computational Linguistics

validate that geometrical models (particularly latent semantic analysis) can produce near human results regarding their propositions on how to account written learner evidence for prior learning and positioning these learners to where the best-suiting starting points are. We will show that latent semantic analysis works for positioning and that it can provide effective positioning.

The main focus of this contribution is to investigate whether machine learning proves useful for the positioning classifiers, whether distractors improve results, and what the role of thresholds for the classifiers is.

The rest of this paper is structured as follows. At first, positioning with LSA and related work are explained. This is followed by an outline of our own approach to positioning. Subsequently, a validation experiment for the set of new algorithms is outlined with which new light is shed on the utilisation of LSA for positioning. The results of this experiment are analysed in the following section in oder to, finally, yield conclusions and an outlook.

## 2 Positioning with LSA

According to (Kalz et al., 2007), positioning "is a process that assists learners in finding a starting point and an efficient route through the [learning] network that will foster competence building". Often, the framework within which this competence development takes places is a formal curriculum offered by an educational provider.

Not only when considering a lifelong learner, for whom the borders between formal and informal learning are absolutely permeable, recognition of prior learning turns out to be crucial for positioning: each individual background differs and prior learning needs to be respected or even accredited before taking up new learning activities – especially before enrolling in a curriculum.

Typically, the necessary evidence of prior learning (i.e., traces of activities and their outcomes) are gathered in a learner's portfolio. This portfolio is then analysed to identify both starting points and a first navigation path by mapping evidence onto the development plans available within the learning network.

The educational background represented in the portfolio can be of formal nature (e.g. certified exams) in which case standard admission and exemption procedures may apply. In other cases such standard procedures are not available, therefore assessors need to intellectually evaluate learner knowledge on specific topics. In procedures for accreditation of prior learning (APL), assessors decide whether evidence brought forward may lead to exemptions from one or more courses.

For supporting the positioning process (as e.g. needed for APL) with technology, three different computational classes of approaches can be distinguished: mapping procedures based on the analysis of informal descriptions with textmining technologies, meta-data based positioning, and positioning based on ontology mappings (Kalz et al., 2007). Latent semantic analysis is one of many possible techniques that can be facilitated to support or even partially automate the analysis of informal portfolios.

### 2.1 LSA

LSA is an algorithm applied to approximate the meaning of texts, thereby exposing semantic structure to computation. LSA combines the classical vector-space model with a singular value decomposition (SVD), a two-mode factor analysis. Thus, bag-of-words representations of texts can be mapped into a modified vector space that is assumed to reflect semantic structure.

The basic idea behind LSA is that the collocation of terms of a given document-term-space reflects a higher-order – latent semantic – structure, which is obscured by word usage (e.g. by synonyms or ambiguities). By using conceptual indices that are derived statistically via a truncated SVD, this variability problem is believed to be overcome.

In a typical LSA process, first a document-term matrix is constructed from a given text base of $n$ documents containing $m$ terms. This matrix $M$ of the size $m \times n$ is then resolved by the SVD into the term vector matrix $T$ (constituting the left singular vectors), the document vector matrix $D$ (constituting the right singular vectors) being both orthonormal and the diagonal matrix $S$.

Multiplying the truncated matrices $T_k$, $S_k$ and $D_k$ results in a new matrix $M_k$ (see Figure 1) which is the least-squares best fit approximation of $M$ with $k$ singular values (Berry et al., 1994).

### 2.2 Related Work

LSA has been widely used in learning applications such as automatic assessment of essays, provision

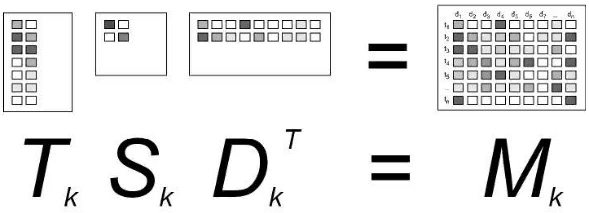

Figure 1: Reconstructing a textmatrix from the lower-order latent-semantic space.

of feedback, and selection of suitable materials according to the learner's degree of expertise in specific domains.

The Intelligent Essay Assessor (IEA) is an example of the first type of applications where the semantic space is build from materials on the topic to be evaluated. In (Foltz et al., 1999) the finding is reported that the IEA rating performance is close to the one of human raters.

In (van Bruggen et al., 2004) authors report that LSA-based positioning requires creating a latent-semantic space from text documents that model learners' and public knowledge on a specific subject. Those texts include written material of learners' own production, materials that the learner has studied and learned in the past, and descriptions of learning activities that the learner has completed in the past. Public knowledge on the specific subject includes educational materials of all kind (e.g. textbooks or articles).

In this case the description of the activity needs to be rich in the sense of terminology related to the domain of application. LSA relies on the use of rich terminology to characterize the meaning.

Following the traditional LSA procedure, the similarity (e.g. cosine) between LSA vector models of the private and public knowledge is then calculated to obtain the learner position with respect to the public knowledge.

## 3 Learning Algorithms for Positioning

In the following, we design an experiment, conduct it, and evaluate the results to shed new light on the use of LSA for positioning.

The basic idea of the experiment is to investigate whether LSA works for advising assessors on acceptance (or rejection) of documents presented by the learner as evidence of previous conceptual knowledge on specific subjects covered by the curriculum. The assessment is in all cases done by comparing a set of learning materials (model solutions plus previously accepted/rejected reference material) to the documents from learners' portfolios using cosines as a proxy for their semantic similarity.

In this comparison, thresholds for the cosine measure's values have to be defined above which two documents are considered to be similar. Depending on how exactly the model solutions and additional reference material are utilised, different assessment algorithms can be developed.

To validate the proposed positioning services elaborated below, we compare the automatic recommendations for each text presented as evidence with expert recommendations over the same text (external validation).

To train the thresholds and as a method for assessing the provided evidence, we propose to use the following five different unsupervised and supervised positioning rules. These configurations differ in the way how their similarity threshold is calculated and against which selection of documents (model solutions and previously expert-evaluated reference material) the 'incoming' documents are compared. We will subsequently run the experiment to investigate their effectiveness and compare the results obtained with them.

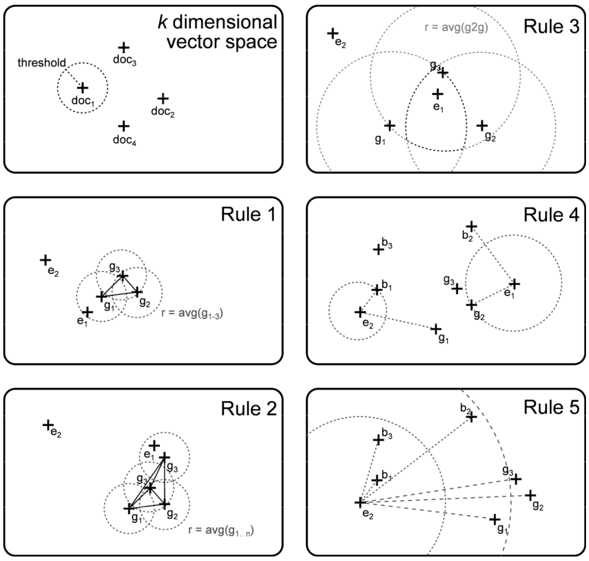

Figure 2: The five rules.

The visualisation in Figure 2 depicts the working principle of the rules described below. In each panel, a vector space is shown. Circles depict radial cosine similarity. The document representatives labelled with $g_n$ are documents with positive evidence ('good' documents), the ones labelled with $b_n$ are those with negative. The test documents carry the labels $e_n$ ('essay').

*Best of Golden*: The threshold is computed by averaging the similarity of all three golden standard essays to each other. The similarity of the investigated essay is compared to the best three golden standard essays (=machine score). If the machine score correlates above the threshold with the human judgement, the test essay is stated correct. This rule assumes that the gold standards have some variation in the correlation among each other and that using the average correlation among the gold standards as a threshold is taking that into account.

*Best of Good*: Best essays of the humanly judged good ones. The assumption behind this is that with more positive examples to evaluate an investigated essay against, the precision of the evaluation should rise. The threshold is the average of the positive evidence essays among each other.

*Average to Good > Average among Good*: Tests if the similarity to the 'good' examples is higher than the average similarity of the humanly judged good ones. Assumption is that the good evidence gathered circumscribes that area in the latent semantic space which is representative of the abstract model solution and that any new essay should be within the boundaries characterised by this positive evidence thus having a higher correlation to the positive examples then they have among each other.

*Best of Good > Best of Bad*: Tests whether the maximum similarity to the good essays is higher than the maximum similarity to bad essays. If a tested essay correlates higher to the best of the good than to the best of the bad, then it is classified as accepted.

*Average of Good > average of Bad*: The same with average of good > average of bad. Assumption behind this is again that both bad and good evidence circumscribe an area and that the incoming essay is in either the one or the other class.

## 4 Corpus and Space Construction

The corpus for building the latent semantic space is constructed with 2/3 German language corpus (newspaper articles) and 1/3 domain-specific (a textbook split into smaller units enhanced by a collection of topic related documents which Google threw up among the first hits). The corpus has a size of 444k words (59.719 terms, 2444 textual units), the mean document length is 181 words with a standard deviation of 156. The term frequencies have a mean of 7.4 with a standard deviation of 120.

The latent semantic space is constructed over this corpus deploying the lsa package for R (Wild, 2008; Wild and Stahl, 2007) using $dimcalc\_share$ as the calculation method to estimate a good number of singular values to be kept and the standard settings of $textmatrix()$ to pre-process the raw texts. The resulting space utilises 534 dimensions.

For the experiment, 94 essays scored by a human evaluator on a scale from 0 to 4 points where used. The essays have a mean document length of 22.75 terms with a standard deviation of 12.41 (about one paragraph).

To estimate the quality of the latent semantic space, the learner writings were folded into the semantic space using $fold\_in()$. Comparing the non-partitioned (i.e. 0 to 4 in steps of .5) human scores with the machine scores (average similarity to the three initial model solutions), a highly significant trend can be seen that is far from being perfect but still only slightly below what two human raters typically show.

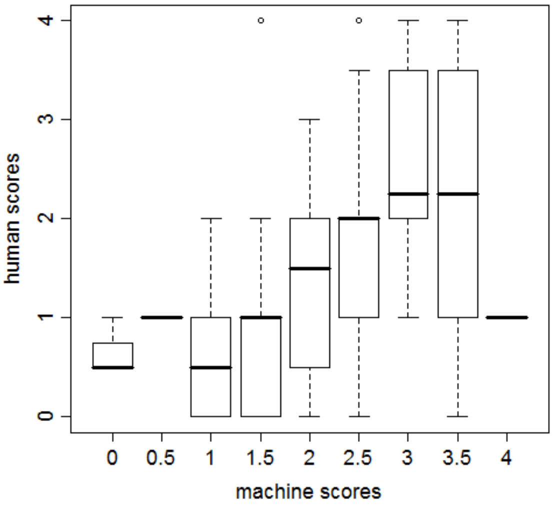

Figure 3: Human vs. Machine Scores.

Figure 3 shows the qualitative human expert judgements versus the machine grade distribution using the non-partitioned human scores (from 0 to 4 points in .5 intervals) against the rounded average cosine similarity to the initial three model solutions. These machine scores are rounded such that they – again – create the same amount of intervals. As can be seen in the figure, the extreme of each score level is displayed in the upper and lower whisker. Additionally, the lower and upper 'hinge' and the median are shown. The overall Spearman's rank correlation of the human versus the (continuous) machine scores suggests a with .51 medium effect being highly significant on a level with the p-value below .001. Comparing this to untrained human raters, who typically correlate around .6, this is in a similar area, though the machine differences can be expected to be different in nature.

A test with 250 singular values was conducted resulting in a considerately lower Spearman correlation of non-partitioned human and machine scores.

Both background and test corpus have deliberately been chosen from a set of nine real life cases to serve as a prototypical example.

For the experiment, the essay collection was split by half into training (46) and test (48) set for the validation. Each set has been partitioned into roughly an equal number of accepted (scores $< 2$, 22 essays in training set, 25 in test) and rejected essays (scores $>= 2$, 24 essays in training, 23 in test). All four subsets, – test and training partitioned into accepted and rejected –, include a similarly big number of texts.

In order to cross validate, the training and test sets were random sampled ten times to get rid of influences on the algorithms from the sort order of the essays. Both test and training sets were folded into the latent semantic space. Then, random sub samples (see below) of the training set were used to train the algorithms, whereas the test set of 48 test essays in each run was deployed to measure precision, recall, and the f-measure to analyse the effectiveness of the rules proposed.

Similarity is used as a proxy within the algorithms to determine whether a student writing should be accepted for this concept or rejected. As similarity measure, the cosine similarity $cosine()$ was used.

In each randomisation loop, the share of accepted and rejected essays to learn from was varied in a second loop of seven iterations: Always half of the training set essays were used and the amount of accepted essays was decreased from 9 to 2 while the number of rejected essays was increased from 2 to 9. This way, the influence of the number of positive (and negative) examples could be investigated.

This mixture of accepted and rejected evidence to learn from was diversified to investigate the influence of learning from changing shares and rising or decreasing numbers of positive and/or negative reference documents – as well as to analyse the influence of recalculated thresholds. While varying these training documents, the human judgements were given to the machine in order to model learning from previous human assessor acceptance and rejection.

## 5 Findings

### 5.1 Precision versus Recall

The experiments where run with the five different algorithms and with the sampling procedures described above. For each experiment precision and recall where measured to find out if an algorithm can learn from previous inputs and if it is better or worse compared to the others.

As mentioned above, the following diagrammes depict from left to right a decreasing number of accepted essays available for training (9 down to 2) while the number of rejected essays made available for training is increased (from 2 to 9).

Rule 1 to 3 do not use these negative samples, rule 1 does not even use the positive samples but just three additional model solutions not contained in the training material of the others. The curves show the average precision, recall, and f-measure[2] of the ten randomisations necessary for the cross validation. The size of the circles along the curves symbolises the share of accepted essays in the training set.

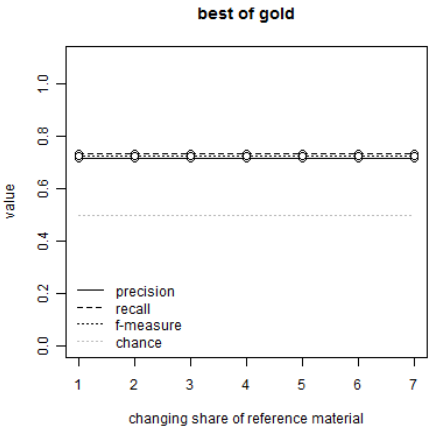

Figure 4: Rule 1: Best of Three Golden

[2] $F = 2 \cdot \frac{precision \cdot recall}{precision + recall}$

Figure 4 shows that recall and precision stay stable as there are no changes to the reference material taken into account: all essays are evaluated using three fixed 'gold standard' texts. This rule serves as a baseline benchmark for the other results.

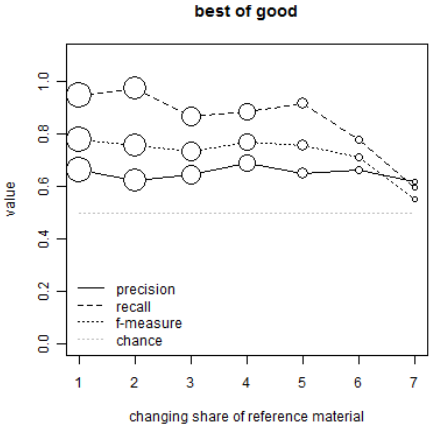

Figure 5: Rule 2: Best of Good

Figure 5 depicts a falling recall when having less positively judged essays in the training sample. In most cases, the recall is visibly higher than in the first rule, 'Best of Gold', especially when given enough good examples to learn from. Precision is rather stable. We interpret that the falling recall can be led back to the problem of too few examples that are then not able to model the target area of the latent semantic space.

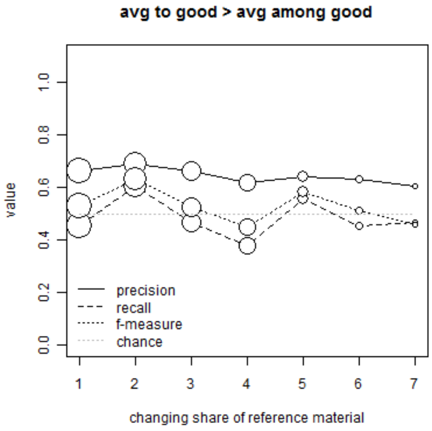

Figure 6: Rule 3: Avg of Good > Avg among Good

Figure 6 displays that the recall worsens and is very volatile[3]. Precision, however, is very stable and slightly higher than in the previous rule, especially with rising numbers of positive examples. It seems that the recall is very dependant on the positive examples whether they are able to characterise representative boundaries: seeing recall change with varying amounts of positive examples, this indicates that the boundaries are not very well chosen. We assume that this is related to containing 'just pass' essays that were scored with 2.0 or 2.5 points and distort the boundaries of the target area in the latent semantic concept space.

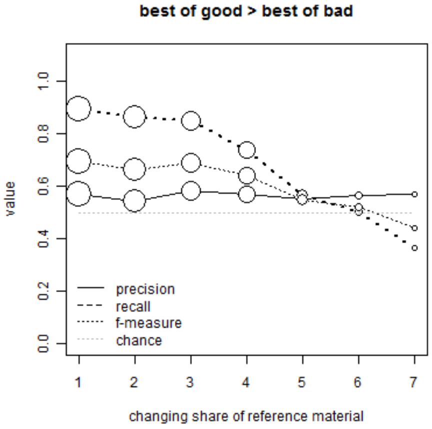

Figure 7: Rule 4: Best of Good > Best of Bad

Figure 7 exhibits a quickly falling recall, though starting on a very high level, whereas precision is relatively stable. Having more negative evidence clearly seems to be counter productive and it seems more important to have positive examples to learn from. We have two explanations for this: First, bad examples scatter across the space and it is likely for a good essay to correlate higher with a bad one when there is only a low number of positive examples. Second, bad essays might contain very few words and thus expose correlation artefacts that would in principle be easy to detect, but not with LSA.

Figure 8 depicts a recall that is generically higher than in the 'Best of Gold' case, while precision is in the same area. Recall seems not to be so stable but does not drop with more bad samples (and less good ones) to learn from such as in the 'Best of Good' case. We interpret that noise can be added to increase recall while still only a low number of positive examples is available to improve it.

[3] We analysed the recall in two more randomisations of the whole experiment; whereas the other rules showed the same results, the recall of this rule was unstable over the test runs, but in tendency lower than in the other rules.

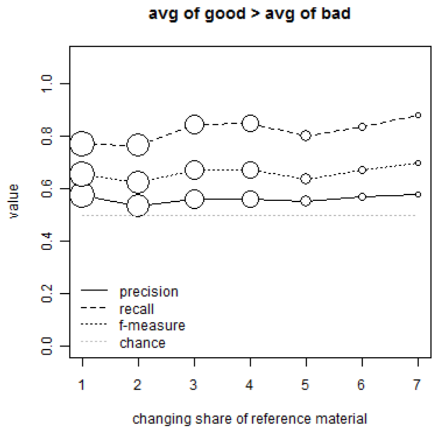

Figure 8: Rule 5: Avg of Good > Avg of Bad

## 5.2 Clustering

To gain further insight about the location of the 94 essays and three gold standards in the higher order latent-semantic space, a simple cluster analysis of their vectors was applied. Therefore, all document-to-document cosine similarities were calculated, filtered by a threshold of .65 to capture only strong associations, and, subsequently, a network plot of this resulting graph was visualised.

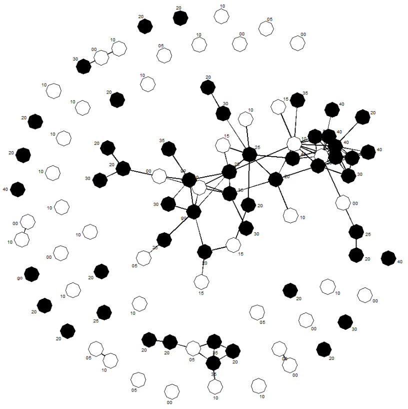

Figure 9: Similarity Network (cos >= .65).

As can be seen in the two charts, the humanly positively judged evidence seems to cluster quite well in the latent-semantic space when visualised as a network plot. Through filtering the document vectors by the vocabulary used only in the accepted, rejected, or both classes, an even clearer picture could be generated, shown in Figure 10.

Both figures clearly depict a big connected component consisting mainly out of accepted essays, whereas the rejected essays mainly spread in the unconnected surrounding. The rejected essays are in general not similar to each other, whereas the accepted samples are.

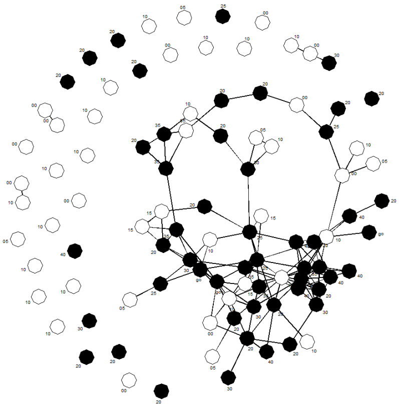

Figure 10: Network with filtered vocabulary.

The second Figure 10 is even more homogeneous than the first due to the use of the restricted vocabulary (i.e. the terms used in all accepted and rejected essays).

# 6 Conclusion and Outlook

Distractors are of low value in the rules tested. It seems that generic noise can be added to keep recall higher when only a low number of positive examples can be utilised. An explanation for this can be found therein that there are always a lot more heterogeneous ways to make an error. Homogeneity can only be assumed for the positive evidence, not for negative evidence.

Noise seems to be useful for the calculation of thresholds. Though it will need further investigation whether our new hypothesis works that bad samples can be virtually anything (that is not good).

Learning helps. The recall was shown to improve in various cases, while precision stayed at the more or less same level as the simple baseline rule. Though the threshold calculation using the difference to good and bad examples seemed to bear the potential of increasing precision.

Thresholds and ways how to calculate them are evidently important. We proposed several well working ways on how to construct thresholds from evidence that extend the state of the art. Thresholds usually vary with changing corpus sizes and the measures proposed can adopt to that.

We plan to investigate the use of support vector machines in the latent semantic space in order to gain more flexible means of characterising the boundaries of the target area representing a concept.

It should be mentioned that this experiment demonstrates that conceptual development can be measured and texts and their similarity can serve as a proxy for that. Of course the experiment we have conducted bears the danger to bring results that are only stable within the topical area chosen.

We were able to demonstrate that textual representations work on a granularity level of around 23 words, i.e. with the typical length of a free text question in an exams.

While additionally using three model solutions or at least two positive samples, we were able to show that using a textbook split into paragraph-sized textual units combined with generic background material, valid classifiers can be built with relative ease. Furthermore, reference material to score against can be collected along the way.

The most prominent open problem is to try and completely get rid of model solutions as reference material and to assess the lower level concepts (terms and term aggregates) directly to further reduce corpus construction and reference material collection. Using clustering techniques, this will mean to identify useful ways for efficient visualisation and analysis.

## 7 Acknowledgements

This work has been financially supported by the European Union under the ICT programme of the 7th Framework Programme in the project 'Language Technology for Lifelong Learning'.

## References

Michael Berry, Susain Dumais, and Gavin O'Brien. 1994. Using linear algebra for intelligent information retrieval. Technical Report CS-94-270, Department of Computer Science, University of Tennessee.

Scott Deerwester, Susan Dumais, Georg W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Peter Foltz, Darrell Laham, and Thomas K. Landauer. 1999. Automated essay scoring: Applications to educational technology. In Collis and Oliver, editors, *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 1999*, pages 939–944, Chesapeake, VA. AACE.

Marco Kalz, Jan Van Bruggen, Ellen Rusmann, Bas Giesbers, and Rob Koper. 2007. Positioning of learners in learning networks with content analysis, metadata and ontologies. *Interactive Learning Environments*, (2):191–200.

Mark K. Smith. 1996. Competence and competency. http://www.infed.org/biblio/b-comp.htm.

Jan van Bruggen, Peter Sloep, Peter van Rosmalen, Francis Brouns, Hubert Vogten, Rob Koper, and Colin Tattersall. 2004. Latent semantic analysis as a tool for learner positioning in learning networks for lifelong learning. *British Journal of Educational Technology*, (6):729–738.

Jan Van Bruggen, Ellen Rusman, Bas Giesbers, and Rob Koper. 2006. Content-based positioning in learning networks. In Kinshuk, Koper, Kommers, Kirschner, Sampson, and Didderen, editors, *Proceedings of the 6th IEEE International Conference on Advanced Learning Technologies*, pages 366–368, Kerkrade, The Netherlands.

Fridolin Wild and Christina Stahl. 2007. Investigating unstructured texts with latent semantic analysis. In Lenz and Decker, editors, *Advances in Data Analysis*, pages 383–390, Berlin. Springer.

Fridolin Wild. 2008. lsa: Latent semantic analysis. r package version 0.61.