eChallenges e-2010 Conference Proceedings
Paul Cunningham and Miriam Cunningham (Eds)
IIMC International Information Management Corporation, 2010
ISBN: 978-1-905824-20-5

Monitoring Conceptual Development with Text Mining Technologies: CONSPECT

Fridolin WILD, Debra HALEY, Katja BÜLOW
The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK
Tel: +44 (0)1908 858524 , Fax: +44 (0)1908 653 169,
Email: f.wild, d.t.haley, k.buelow @ open.ac.uk

Abstract: This paper evaluates CONSPECT, a service that analyses states in a learner's conceptual development. It combines two technologies – Latent Semantic Analysis to analyse text and Network Analysis (NA) to provide visualisations – into a technique called Meaningful Interaction Analysis (MIA). CONSPECT was designed to help both online learners and their tutors monitor their conceptual development. This paper reports on the validation experiments undertaken to determine how well LSA matches first year medical students in clustering concepts and in annotating text. The validation used several techniques, including card sorting and Likert scales. CONSPECT produces almost 'peer' quality results and what remains to be tested is whether it improves with more advanced learners. One of the experiments showed an average 0.7 correlation between humans and CONSPECT.

1. Introduction – Issues to be addressed

This paper describes and evaluates CONSPECT (from concept inspection), an application that analyses states in a learner's conceptual development. It was designed to help online learners monitor their conceptual development and also to help reduce the workload of tutors monitoring a learner's conceptual development.

CONSPECT combines two technologies – Latent Semantic Analysis (LSA) and Network Analysis (NA) into a technique called Meaningful Interaction Analysis (MIA). LSA analyses the language and NA provides visualisations of the information calculated by LSA.

This paper reports on the validation activities undertaken to show how well LSA matches first year medical students in 1) grouping similar concepts and 2) annotating text.

1.1 Theoretical Underpinning

This subsection mentions two related Computational Linguistic theories that support the approach taken in CONSPECT: Fauconnier's Mental Spaces Theory and Conceptual Blending Theory [1]. These theories hold that the meaning of a sentence cannot be determined without considering the context. Meaning construction results from the development of mental spaces, also known as conceptual structures [2], and the mapping between these spaces.

Mental spaces and their relationships are what LSA tries to quantify. LSA uses words in their contexts to calculate semantic similarity. This use of context is consistent with Fauconnier's claim that context is crucial to construct meaning.

Some researchers use network analysis to analyse conceptual structures. Schvaneveldt et al [3], Goldsmith et al [4] and Clariana & Wallace [5] are among the researchers who use a particular class of network called Pathfinder, which are derived from proximity data [3]. These researchers assume that "concepts and their relationships can be represented by a structure consisting of nodes (concepts) and links (relations)." The strength of the

relationships can be measured by the link weights. The networks of novices and experts are compared to gauge the learning of the novices.

Pathfinder techniques require the creation of proximity matrices by association, or relationship testing. LSA, on the other hand, requires no such explicit proximity judgments. It uses textual passages to compute automatically a proximity matrix. Thus LSA requires less human effort than these other techniques.

1.2 Latent Semantic Analysis

The subsection briefly explains LSA, a statistical natural language processing technique whose purpose is to analyse text. The interested reader can learn more by visiting a web site containing almost 100 papers [6].

LSA is similar to the vector space model [7], which uses a large corpus related to the knowledge domain of interest and creates a term/document matrix whose entries are the number of times each term appears in each document. The LSA innovation is to transform the matrix using singular value decomposition (SVD) and reduce the number of dimensions of the singular value matrix produced by SVD, thus reducing noise due to chance and idiosyncratic word choice. The result provides information about the concepts in the documents as well as numbers that quantify the semantic similarity between terms and documents, terms and terms, and documents and documents.

1.3 Objectives of Paper

The objectives of this paper are as follows:

- introduce CONSPECT and describe how it can help monitor a learner's conceptual development
- provide a theoretical basis for our method of measuring a learner's conceptual development
- discuss the technologies employed by CONSPECT
- describe the annotation experiment to validate the accuracy of LSA
- describe the clustering experiment to validate the accuracy of LSA
- discuss the results of the experiments

2. Technology Description

2.1 The User Point of View

CONSPECT is a web based service that uses widgets in its user interface. After logging in to the service using openID, the learner is shown a list of RSS feeds (usually blogs or learning diaries) and conceptograms (graph-like visualisations of the output of the LSA processing). The user can add a new feed, view a conceptogram, or combine two conceptograms. A simple conceptogram shows the concepts written about in the feed. A combined conceptogram compares the concepts of two entities; for example, if the learner combines a conceptogram showing a course's intended learning outcomes with his personal conceptogram, he can see which of the intended outcomes he has covered, which he has not covered, and which concepts he has written about that are not part of the intended learning outcomes.

Similarly, a tutor can monitor the progress of her learners by inspecting a single learner's conceptogram and can combine the conceptograms of two students to compare one with the other. Other possibilities are to compare one learner's conceptograms over time and to compare a learner's conceptogram to the group's emergent reference model.

2.2 The Background Processing

A great deal of processing takes place before the user can see a conceptogram. First, an LSA semantic space must be created from a knowledge domain-specific training corpus. Next, the feed is converted to and folded in to the original semantic space. The concepts are filtered so that the similarity is assigned zero for all similarities less than 0.7 or one for all similarities greater than or equal to 0.7. Next, certain ideas from network analysis (e.g., degree centrality, closeness [8]) are used to create graphs. Finally, the graphs are displayed using a force-directed layout technique [9].

3. Methodology

Eighteen first year medical students participated in the experiments; by chance, half were female and half were male. They received a £10 book voucher for their participation. These students were a target audience for CONSPECT.

3.1 Experiment 1: Clustering

Experiment 1 examined whether humans cluster concepts in the same way as does CONSPECT. It was a type of card-sorting evaluation.

Preparation: CONSPECT generated a list of about 50 concepts for five documents from authentic postings about "safe prescribing". The concepts were printed on a set of cards; this yielded five sets of about 50 cards in each set for each participant.

Procedure: The researcher gave sets of cards to the participants and asked them to arrange the cards into groups so that each group contained strongly associated concepts. The participants decided on the number of categories but it had to be more than one and less than the number of cards in the set, that is, there had to be more than one category and each category had to have more than one card. The experimenter then recorded the concepts and the categories chosen by the participant.

Analysis: The analysis provided information on how closely humans agree with CONSPECT's concept classifications. This analysis was undertaken in two ways. First, the researcher used Diebel et al's [10] metric of edit distances. The analysis showed that the 18 human participants were about 10% better than was CONSPECT in clustering concepts. See [11] for details.

3.2 Experiment 2: Text Annotation

This subsection discusses a second type of analysis of the card sort data, which looked at co-occurrence matrices. This type of matrix is of size n x n where n is the number of concepts to group. Each entry (i,j) ranges from zero to the number of participants. It shows how many of the participants clustered term i and term j in the same category.

Experiment 2 looked at whether humans agreed with the descriptors that CONSPECT assigned to a text.

Preparation: CONSPECT generated ten descriptors for each of five texts obtained from postings about safe prescribing and five "distracter" descriptors. The first ten descriptors were those that had the highest similarity to the texts. The distracter descriptors were chosen randomly from the related descriptors. These fifteen descriptors were printed in alphabetical order on a sheet of paper along with the text of the posting.

Procedure: Each participant was given five sheets of paper, one for each test and were asked to rank each descriptor on a Likert scale of 1 to 5 based on whether they thought the concept was descriptive of the post.

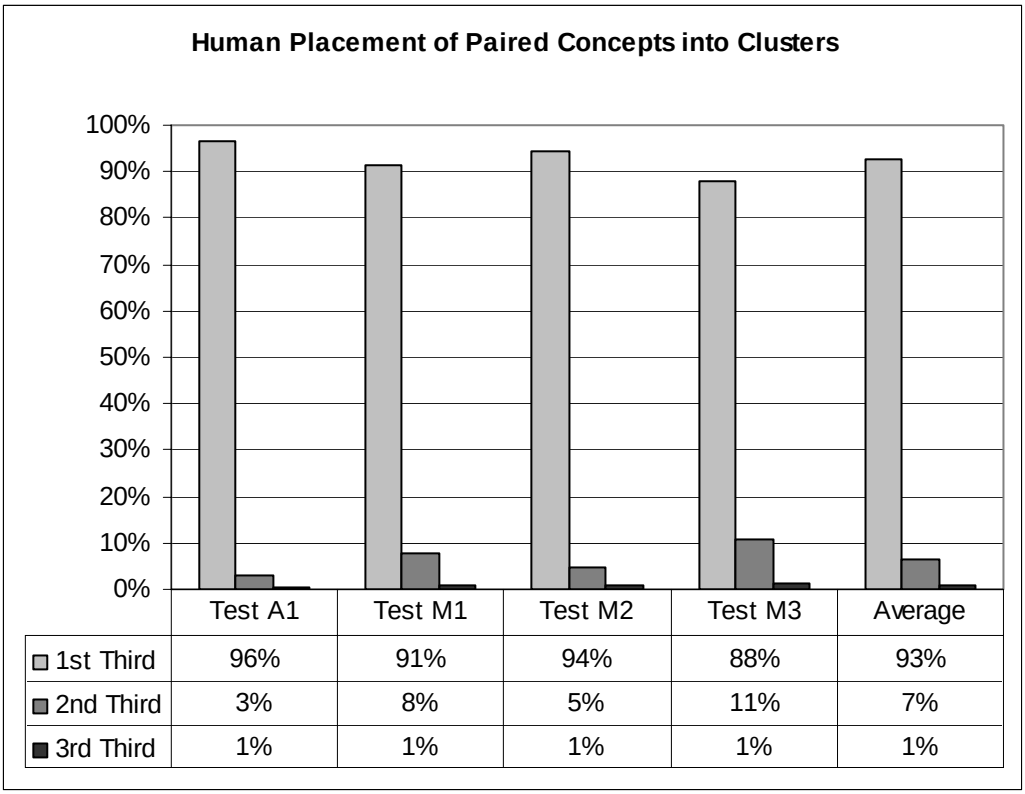
Analysis: Two techniques were used to analyse the text annotation data. First, a picture of the spread of annotations was created using bar charts. Second, inter-rater reliability figures were calculated with kappa.

4. Discussion

This section presents the results of the validation experiments.

4.1 Experiment 1 - Clustering

Figure 1 shows the spread of data from the co-occurrence matrices. The bar chart shows a noted similarity between the four postings. On average, the vast majority of the paired concepts were in the bottom third, that is, 93% of the pairs were put in the same group by from 0 to 6 participants. Just 7% of the pairs had between 7 and 12 participants placing them in the same cluster. A tiny number, just 1% of the pairs, were placed in the same cluster by more than 12 of the participants. These groups are referred to as the first, second, and third "thirds".



	Test A1	Test M1	Test M2	Test M3	Average
1st Third	96%	91%	94%	88%	93%
2nd Third	3%	8%	5%	11%	7%
3rd Third	1%	1%	1%	1%	1%

Figure 1Human Placement of Concepts in Clusters

Figure 2 compares the clustering done by the participants with that done by CONSPECT. The data come from the clustering done by the participants, grouped into thirds. The top third of the groupings comprise the co-occurrence figures ranging from 13 through 18 compared with the LSA cosine similarity figures for the same pairs of terms. The correlation between these two sets of figures was calculated using Spearman's rho. Only this subset of pairs was considered due to time limitations. If all the pairs were to be analysed, over four thousand calculations would need to be done. (The evaluation used Excel, which does not contain a Spearman's rho function. Thus, the calculations were done by hand. Software would be needed to automate the calculations.)

The bar chart shows the correlations for those pairs of clustered concepts in the top third. Two of the texts showed good correlation (0.7), one showed excellent correlation

(0.9) and one showed moderate correlation (0.4). The correlations averaged 0.7 – a good correlation.

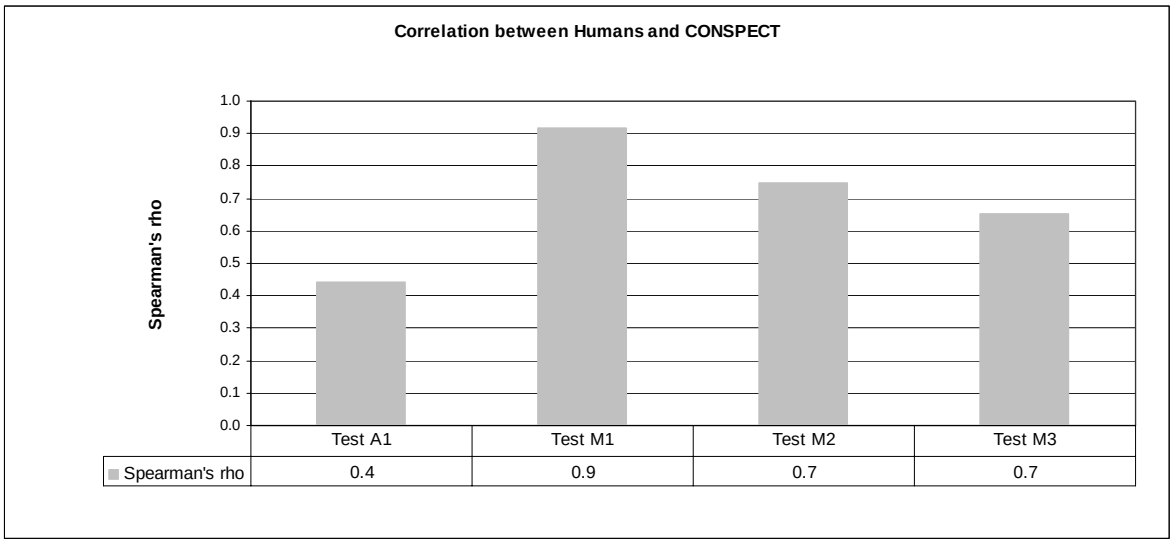


Figure 2 Correlations between Humans and CONSPECT

4.2 Experiment 2 – Text Annotation

The text annotation data was analysed by the free marginal kappa figure [12, 13], a type of inter-rater reliability statistic that is applicable when the raters are not constrained by the number of entries per category. The data come from the Likert selections, that is, the judgments of the participants as to as closely a concept described a text`.

The first type of analysis is given in Figure 3 and Figure 4, which show stacked bar charts for non-conflated and conflated categories, respectively. From the bottom, the Likert categories were “not at all descriptive”, “not very descriptive”, “neutral”, “somewhat descriptive” and “very descriptive”. When distracters are used, more descriptors fall into the bottom two categories – not surprising since distracters were randomly selected and not chosen for their high similarity to the text. Figure 4 is a bit easier to interpret – the two bottom categories were conflated, as were the two top categories.

Tables 1 and 2 below show a different type of analysis. Table 1 shows the results for five categories; Table 2 shows the results for 3 categories (i.e., categories 1 and 2 were conflated, as were categories 4 and 5). Each table gives kappa inter-rater reliability figures for three sets of data: all 15 terms (descriptors plus distracters), for ten descriptors, and finally for just the five distracters. Table 1 shows the highest agreement occurs when only the distracters are considered and the lowest agreement when the distracters are removed. Table 2 shows a similar pattern when conflated categories are examined. In each case (i.e. conflated and non-conflated categories) the reliability figure is lower than the accepted threshold of 0.7 [12] except when just the distracters were examined.

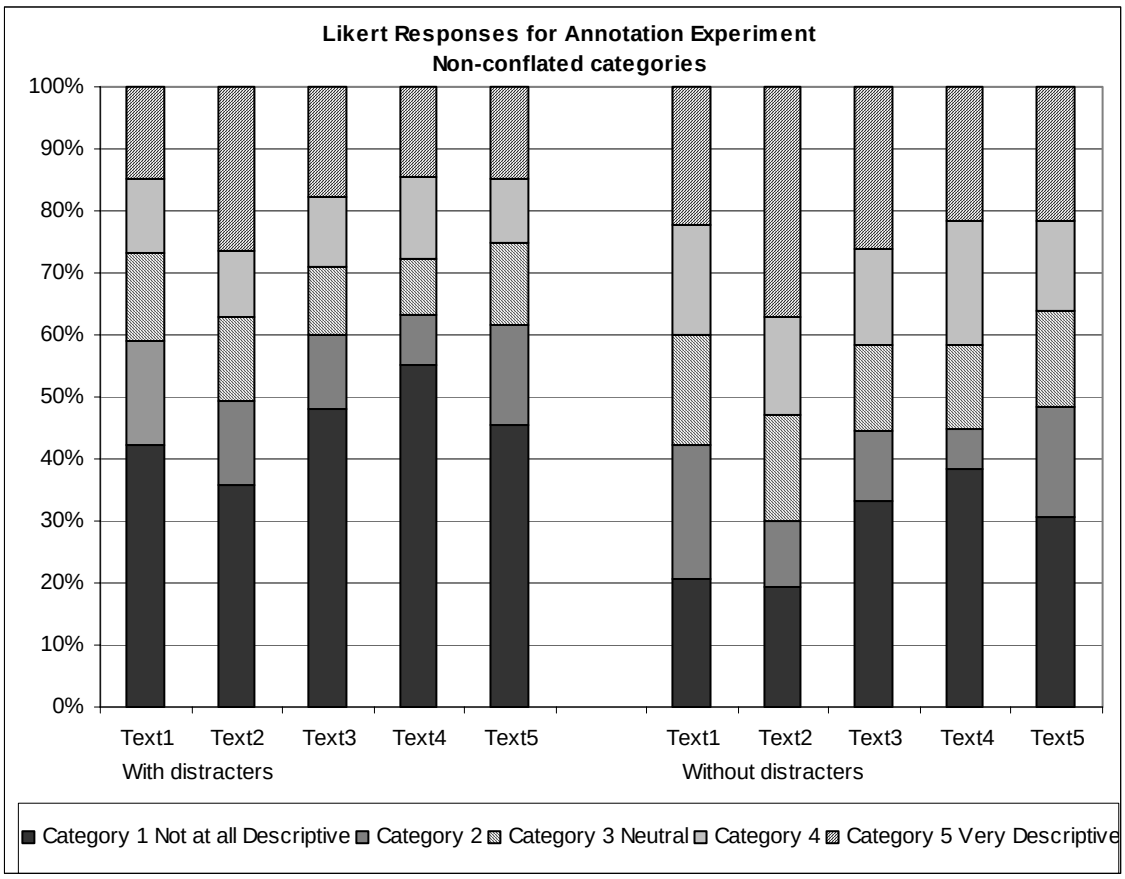


Figure 3 Likert Responses for Annotation Experiment - Non-Conflated Categories

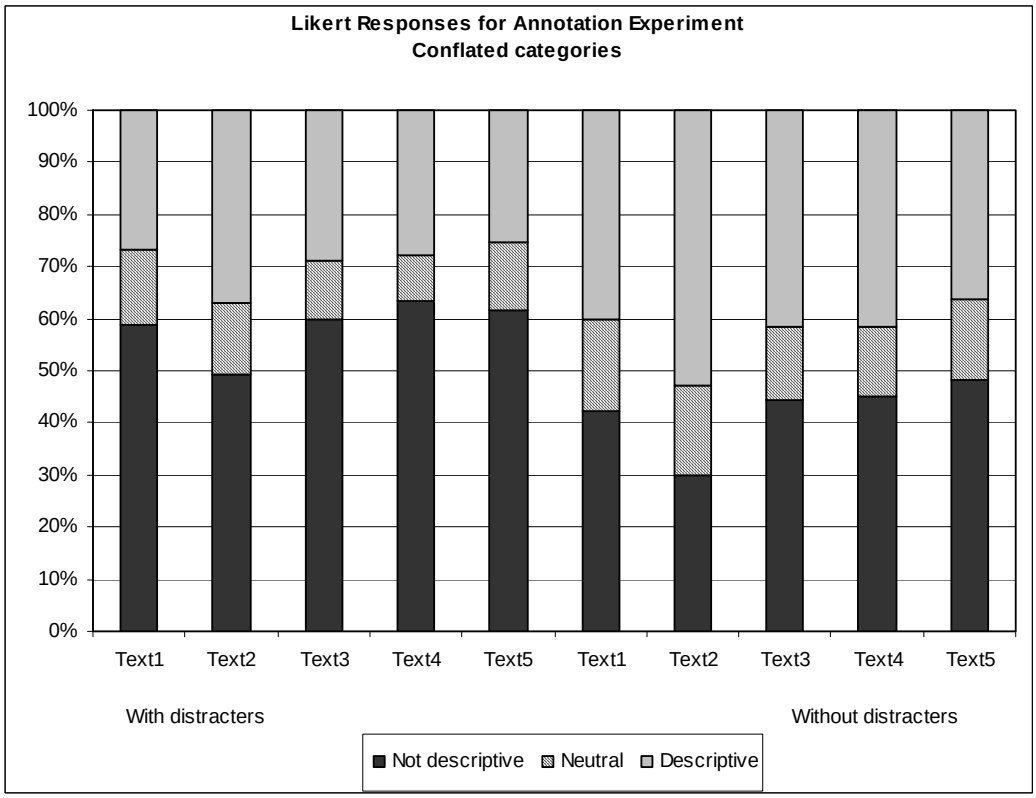


Figure 4 Likert Responses for Annotation Experiment - Conflated Categories

Table 1 Inter-Rater Agreement Between Humans and CONSPECT

	free marginal kappa	with distracters removed	only distracters
Text 1	0.4	0.2	0.7
Text 2	0.4	0.3	0.5
Text 3	0.4	0.3	0.5
Text 4	0.4	0.3	0.8
Text 5	0.3	0.2	0.5
Average	0.4	0.3	0.6

Table 2 Inter-Rater Agreement with Categories 1 and 2 and 4 and 5 Conflated

	free marginal kappa	no distracters	only distracters
Text 1	0.5	0.4	0.8
Text 2	0.5	0.4	0.7
Text 3	0.6	0.4	0.8
Text 4	0.6	0.4	1.0
Text 5	0.5	0.4	0.7
Average	0.5	0.4	0.8

5. Industrial Significance and Eventual Benefits

CONSPECT has benefits for any institution needing to evaluate the conceptual development of current or potential employees or students. Its purpose is not to provide summative results but rather formative assistance. It provides an automatic way to assess the knowledge of learners – both by the learners themselves and anyone wishing to evaluate an individual.

There are two prerequisites for using CONSPECT. One need is to locate/create an acceptable training corpus. Both the language and domain of the corpus need to be tailored to the application. The second need is to acquire a written text from the individual.

One problem noted in evaluation experiments is difficulty in interpreting the conceptograms. Improvements to these visualizations, including purely textual information, are in progress and will be tested with both British medical students and Dutch psychology students.

In addition to improving the user interface, further experiments to measure the accuracy of the results are ongoing, which may result in changes to the underlying algorithms. The development will be completed by the end of the year. CONSPECT is part of the Language Technologies for Lifelong Learning (LTfLL) project (http://www.ltfll-project.org/), which is responsible for its dissemination.

6. Conclusions

This paper described the activities undertaken to validate CONSPECT. Two experiments were conducted – card sorting and text annotation. All of the analyses of these two experiments show that CONSPECT is better at identifying terms that do *not* describe a text well than those that *do* describe a text well. However, good correlation (0.7 on average) was found between humans and CONSPECT in the clustering experiment.

The kappa results from the text annotation experiment show that humans do not have very good agreement with each other; the average is only 0.4, where 0.7 is considered good. As to be expected, both humans and CONSPECT do slightly better when the lower two and upper two categories are conflated. Humans judge that the terms chosen by CONSPECT agree with a text at an average kappa of either 0.4 or 0.5. When only distracters are analysed, humans judge that the terms not chosen by CONSPECT describe a text with a kappa of either 0.6 or 0.8, depending on whether or not the categories are conflated. CONSPECT can be used to eliminate those terms that do not apply and since the human agreement was so low, it remains to be tested, whether it works better with more advanced learners or communities of practice that share a frame of reference.

It would be instructive to repeat these experiments in another language and in another domain. In principle, nothing about the way that LSA works is language or domain

dependent (aside from the training corpus). We are in the process of repeating these experiments using psychology students in Dutch.

Future work is planned to further analyse all of the co-occurrence data, rather than just a sampling of the top and bottom categories. This analysis requires software to be written due to the large number of calculations needed.

In addition, several improvements to CONSPECT are planned, including increasing the number of iterations in the k-means clustering algorithm or finding a better algorithm. Another change is to vary the thresholds. Since the algorithms used cannot be evaluated isolated from their underlying corpus and latent semantic space, further insights are expected with new and possibly better corpora.

Acknowledgments

CONSPECT was developed as a part of the Language Technologies for Life Long Learning (LTfLL) project (see http://ltfll-project.org/).The LTfLL project is funded by the European Union under the ICT programme of the 7th Framework Programme (Contract number: 212578). We would like to thank the University of Manchester, specifically Alisdair Smithies, for help and support in this investigation.

References

[1] Evans, V. and M. Green, *Cognitive Linguistics, An Introduction*. 2006, Edinburgh: Edinburgh University Press.

[2] Saeed, J.I., *Semantics*. 3rd ed. 2009: Wiley-Blackwell.

[3] Schvaneveldt, R.W., et al., *Network Structures in Proximity Data.* The Psychology of Learning and Motivation, 1989. **25**: p. 249-284.

[4] Goldsmith, T.E., P.J. Johnson, and W.H. Acton, *Assessing Structural Knowledge.* Journal of Educational Psychology, 1991. **83**(1): p. 88-96.

[5] Clariana, R.B. and P. Wallace, *A Computer-based Approach for Deriving and Measuring Individual and Team Knowledge Structure from Essay Questions.* Journal of Education Computing Research, 2007. **37**(3): p. 211-227.

[6] Lemaire, B. *Readings in Latent Semantic Analysis*. Available from: http://membres-timc.imag.fr/Benoit.Lemaire/lsa.html.

[7] Salton, G., A. Wong, and C.S. Yang, *A vector space model for automatic indexing.* Communications of the ACM, 1975. **18**(11): p. 613-620.

[8] Brandes, U. and T. Erlebach, *Network Analysis: Methodological Foundations*. 2005, Berlin, Heidelberg: Springer-Verlag.

[9] Fruchterman, T. and E. Reingold, *Graph Drawing by Force-directed Placement.* Software - Practice and Experience, 1991. **21**(11): p. 1129-1164.

[10] Deibel, K., R. Anderson, and R. Anderson, *Using Edit Distance to Analyze Card Sorts.* Expert Systems, 2005. **22**(3): p. 129-138.

[11] Wild, F., D. Haley, and K. Bulow. *CONSPECT: Monitoring Conceptual Development.* in *ICWL*. submitted. Shanghai.

[12] Randolph, J.J., *Free-marginal multirater kappa: An alternative to Fleiss' fixed-marginal multirater kappa*, in *Joensuu University Learning and Instruction Symposium*. 2005: Joensuu, Finland.

[13] Randolph, J.J. *Online Kappa Calculator*. 2005. Last accessed 7 May 2010 Available from: http://justusrandolph.net/kappa/.