
Using Language Technologies for Monitoring Conceptual Development

---

Fridolin WILD[1], Debra HALEY[1], Katja BÜLOW[1]
[1]*The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK*
Tel: +44 (0)1908 858 524 , Fax: +44 (0)1908 653 169
Email: f.wild, d.t.haley, k.buelow @ open.ac.uk

Abstract

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 and their tutors monitor 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 meaning in the textual digital traces left behind by learners in their learning journey; NA provides the analytic instrument to investigate (visually) the semantic structures identified by LSA.

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

1. Theoretical Justification

This section mentions two related Cognitive Linguistic theories that support the approach taken in CONSPECT: Fauconnier's Mental Spaces Theory and Conceptual Blending Theory (Evans and Green 2006). 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 (Saeed 2009), 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 associative closeness. This use of context is consistent with Fauconnier's claim that context is crucial to construct meaning.

Various researchers use network analysis to analyse conceptual structures: Schvaneveldt et al (1989), Goldsmith et al (1991) and Clariana & Wallace (2007) are among the researchers who use a particular class of networks called Pathfinder, which are derived from proximity data (Schvaneveldt, Durso et al. 1989). 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.

1.1 Latent Semantic Analysis

The subsection briefly explains LSA, a statistical natural language processing technique whose purpose is to analyse text. For a comprehensive introduction to LSA, see Landauer et al (2007). LSA was chosen as the analysis technique due to the extensive literature reporting positive results.

LSA is similar to the vector space model (Salton, Wong et al. 1975), 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 reflect associative closeness between terms and documents, terms and terms, and documents and documents.

2. Technology Description

2.1 Overview

CONSPECT, a web-based, widgetised service accepts RSS feeds as input, processes the data using LSA and network analysis, and outputs results visually in the form of conceptograms or non-graphically as lists.

The RSS feeds are in the form of text from blog posts or learning diaries. These posts and diaries are assumed to be a normal part of a learner's course work or a tutor's preparation. It would be possible to write the code to allow Word documents as input; however, time constraints have not allowed this enhancement. The decision was made to use feeds from blogs rather than Word documents because learner reflection in the form of blogs or online learning diaries has become common.

CONSPECT allows the user to include all or some of the blog posts in the feed, thus providing more flexibility.

An example of a conceptogram is shown in Figure 1 and is described below.

2.2 The user point of view

After logging in using openID, the learner is shown a list of existing RSS feeds and conceptual graphs, called conceptograms. Each graph can be inspected in a visualisation using force-directed layouts (conceptograms) of the output of the LSA processing. The user can add a new feed, view a conceptogram, or create an agreement plot between two distinct conceptual representations, i.e., a combined conceptogram.

A single conceptogram shows the concepts written about in the feed. Various types of single conceptograms can be produced using appropriate feeds. An individual student conceptogram would show which concepts the student has written about in the blog. A single conceptogram showing the course's intended learning outcomes (ILO) would come from a tutor-provided blog post giving the ILO. A combined conceptogram compares the concepts of two graphs; for example, if the learner compares a conceptogram showing a course's intended learning outcomes with the conceptogram of his personal learning history, he can see which of the intended outcomes he has covered, which he has not covered, and which concepts he has written about that go beyond the intended learning outcomes.

Similarly, a tutor can monitor the progress of her learners. By aggregating conceptual graphs of several learners, reference models can be constructed. 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 (created by combining individual student conceptograms covering a particular time frame). Figure 1 shows a combined conceptogram that compares the concepts of two learners. The real version has three colours - one colour would show the concepts discussed by both students, one colour would show concepts discussed by Student 1 but not Student 2 and similarly for the third colour. Figure 1 shows that both students discussed *type*, *diabet*, *insulin*, and *time*. (The words are stemmed.) Student 1 wrote about *experi*, *parent*, *child*, and *matern* among other concepts, none of which Student 2 covered. Some of the concepts that Student 2 wrote about that were neglected by Student 1 were *nsaid*, *treatment*, *heart*, *obes*, and *glucos*.

2.3 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, a feed is used like a typical LSA document; it is converted to and folded in to the original semantic space. The concepts are filtered so that those closeness relations with cosine proximities below the similarity threshold of 0.7 are assigned zero and eliminated from the latent semantic network. Next, ideas from network analysis (Brandes and Erlebach 2005)) are used to identify structural properties of the graph. Several rendering techniques rerepresent the conceptual graph structure to the end-user. Finally, the graphs are displayed using a force-directed layout technique (Fruchterman and Reingold 1991) which is deployed to create 2D representations.

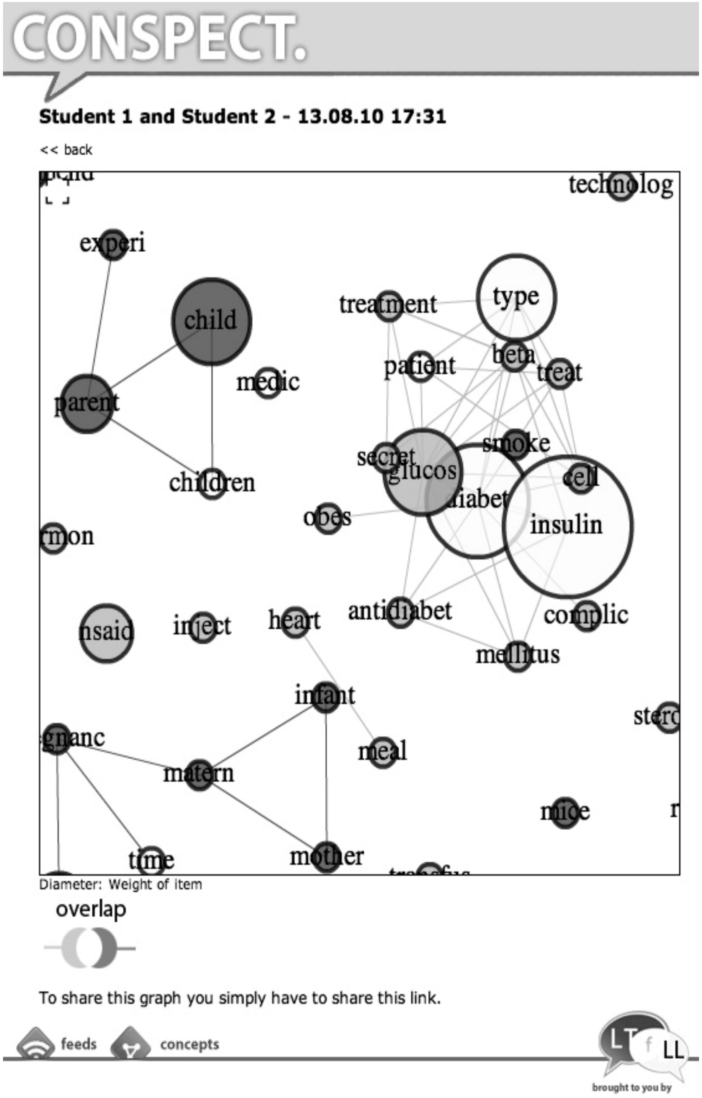

Figure 1. A combined conceptogram showing, overlapping and non-overlapping concepts converted to gray scale for publishing

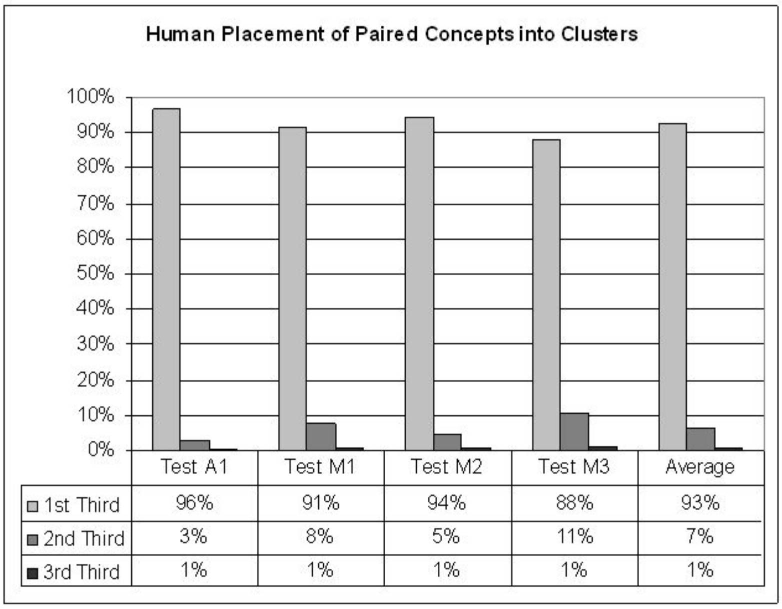

Figure 2. Human Placement of Concepts in Clusters

Table 1 Card Sort Result

		Min	Quartile 1	Quartile 2	Quartile 3	Max	#comparisons
Test A1	Average	21	26.6	28.7	29.8	34	153
CONSPECT		30	31.25	33	34	35	18
Test M1	Average	18	22.6	24.7	26.4	31	153
CONSPECT		24	27	28	29	31	18
Test M2	Average	21	30.9	32.7	34.5	40	153
CONSPECT		31	33.3	35.5	36	37	18
Test M3	Average	20	28.7	31.0	33.1	38	153
CONSPECT		30	31	32.5	35	37	18

Table 2. Edit Distance Information

Sort	**Difference**	%diff	#cards	%of cards	words in posts
3	**2.26**	6.5%	52	67%	1117
4	**2.33**	7.0%	51	65%	533
2	**3.2**	11.5%	43	65%	557
1	**4.5**	13.6%	48	68%	228
average		9.7%			

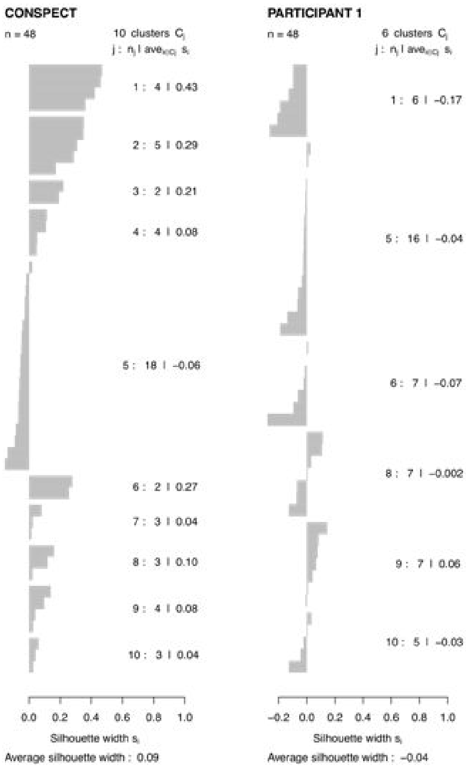

Figure 3. Silhouette plots

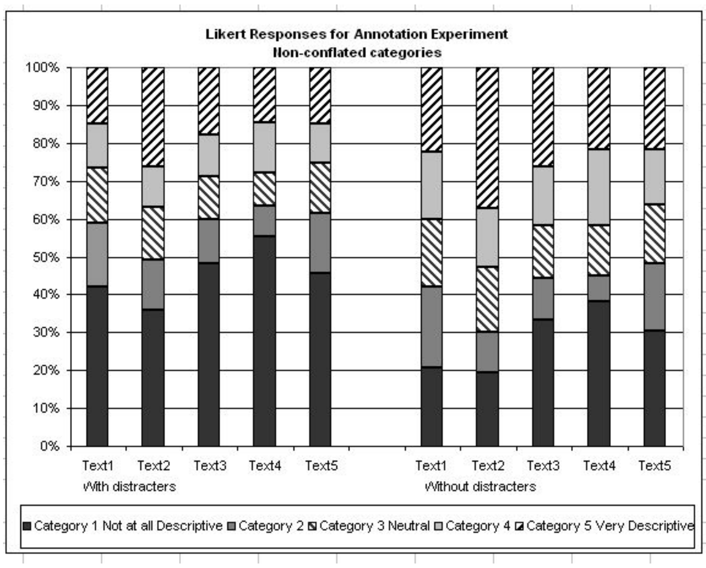

Figure 4. Non-Conflated Likert Responses for Annot'n Exper.

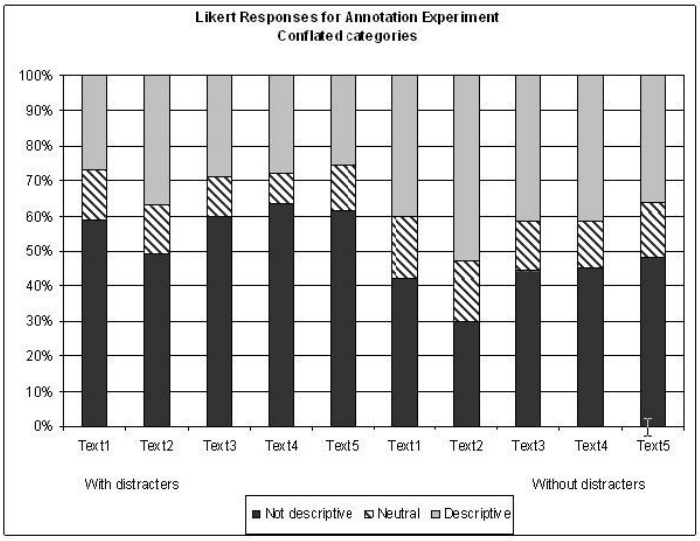

Figure 5. Conflated Likert Responses for Annotation. Experiment

3. Evaluation: Verification

Two types of evaluation of CONSPECT were carried out: verification and validation. Verification tries to determine if the system was built correctly while validation looks at whether the right system was built. The validation results are discussed elsewhere. This section describes the two verification experiments that were conducted: cluster analysis and text annotation. Eighteen first year medical students participated in both experiments; by chance, half were female and half were male. They ranged in age from about eighteen to twenty-two. Each student received a £10 book voucher for participating.

3.1 Experiment 1: clustering

The accuracy of CONSPECT was verified in Experiment 1, which examined whether humans cluster concepts in the same way as does CONSPECT. It was a type of card-sorting (Rugg and McGeorge 1997), a technique often used by web designers but used here in a more unusual way. Card sorts allow a researcher to view a participant's mental model of the words on the cards, which is exactly what was wanted. There is a rich literature on how to conduct card sorts (Rugg and McGeorge 1997; Upchurch, Rugg et al. 2001) particularly relating to web page design, which is characterised by a relatively small number of words. This kind of data is often interpreted qualitatively. It is harder to find advice on how to interpret card sorts with a large number of words. Diebel (2005) encountered just such a problem and developed the concept of edit distance of card sorts to analyze her data. The edit distance is the number of cards from one card sort that must be moved from one pile to another in order to match another card sort.

3.1.1 Methodology

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 semantically similar 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.

3.1.2 Discussion

The analysis provided information on how closely humans agree with CONSPECT's concept classifications. (The classes arise from the LSA cosine similarity measures.) This analysis was undertaken in three ways.

First, the researcher used co-occurrence matrices. Figure 2 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”.

The second analysis used Diebel et al’s (2005) metric of edit distances. The analysis showed that the 18 human participants were about 10% better than was CONSPECT in clustering concepts. Table 1 shows the results of four card sorts, each sort conducted by 18 participants. The table reports on the minimum, maximum, and 1st, 2nd, and 3rd quartile edit distances found by the UW Card Sort Analyzer [2010] for the participants. The lines labelled CONSPECT show the same information when it was compared with the 18 participants. (CONSPECT’s sorting data are the clusters calculated by LSA.)

The data in the table were calculated as follows. For each test, the edit distance was calculated between Participant 1 and each of the other 17 participants. Next, Participant 2 is compared with participants 3 through 18, and so on until all possible pairs are calculated resulting in 153 comparisons. For the 153 comparisons of Sort A1, the minimum distance was 21, the maximum distance was 34 and the averages of the quartiles are 26.6, 28.7, and 29.8.

By looking at the edit distance information, one can compare how CONSPECT performs in relation to the human participants. For the min, max, and average quartile edit distances, the CONSPECT figures are larger in each case.

Table 2 shows the results of an attempt to further understand the card sorting data from Table 1. An interesting question was whether or not the edit distances were dependent on a particular variable. The first column, *Sort*, is the number of the sort. The second column, *difference*, was calculated by subtracting the average of the means. The third column, *%diff*, is the difference divided by average of the means. The fourth column, *#cards*, is the number of cards sorted by the participants. The fifth column, *%of cards* is the #cards divided by the average of the means, so this indicates, for example, that out of a total of 52 cards for Sort 3, 67% of them had to be moved (i.e., the edit distance) to achieve identical sort piles. Finally, the last column, *words in posts*, is the number of words captured in the blog and used to extract the concepts to be sorted. There is no clear relationship among these variables. Therefore, one cannot say that shorter posts result in larger edit distances, for example.

The third column indicates how much larger (as a percentage) the edit distances were for CONSPECT than for the human participants. These figures range from 6rfr .5% to 13.6% with a mean of 9.7%. This analysis suggests that CONSPECT has an edit distance of about 10% larger than the human participants.

The third type of analysis created silhouette plots that showed how well CONSPECT created its clusters. Figure 3 shows the plots. The average silhouette width is .09 for CONSPECT and between -.1 and -.03 for the participants. This means that although the machine was clustering slightly better than the participants, the clusters chosen in all 19 cases were not necessarily very discriminate (but also definitely not bad, which would have been reflected in an average silhouette width of -1.0).

3.2 Experiment 2: text annotation

Experiment 2 looked at whether humans agreed with the descriptors that CONSPECT assigned to a text, i.e., it compared the annotations that humans made to a text with CONSPECT's annotations. The same participants were used as were used in the card sorting experiment.

3.2.1 Methodology

Preparation: CONSPECT generated ten descriptors (those with the highest similarity) for each of five texts obtained from postings about safe prescribing; additionally, five "distracters" were chosen randomly from the available vocabulary. 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.

3.2.2 Discussion

Three techniques were used to analyse the text annotation data.

The first and second techniques to analyse the text annotation data used the free marginal kappa figure (Randolph 2005; Randolph 2005) 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 how closely a concept described a text.

Figure 4 and Figure 5, 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 5 is a bit easier to interpret – the two bottom categories were conflated, as were the two top categories.

Tables 3 and 4 below show a different type of analysis. Table 3 shows the results for five categories; Table 4 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 3. 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 4. Inter-rater agreem't with cat. 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

Table 3 shows the highest agreement occurs when only the distracters are considered and the lowest agreement when the distracters are removed. Table 4 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 (Randolph 2005) except when just the distracters were examined.

Finally, the agreement between humans and CONSPECT was evaluated. More specifically, the percentage of judgements where the humans gave a lower Likert rating for a distracter compared to each descriptor was calculated. Figure 6 shows that the average agreement was 89%. This finding, along with that shown In Table 3 and Table 4, leads to the conclusion that CONSPECT is better at identifying whether a concept *is not* descriptive than it is at deciding whether a concept *is* descriptive.

4. Conclusion

The overall conclusion, based on the results of these several analyses, is that CONSPECT shows enough agreement with humans that it can serve as a valuable tool for monitoring conceptual development. However, further investigations are planned to improve the results, such as finding a better clustering algorithm and adjusting the threshold. In addition, the verification experiments will be repeated with Dutch Psychology students. This will provide very interesting data about how well CONSPECT works with a different language in a different knowledge domain.

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 Univer-

sity of Manchester, specifically Alisdair Smithies, for help and support in this investigation.

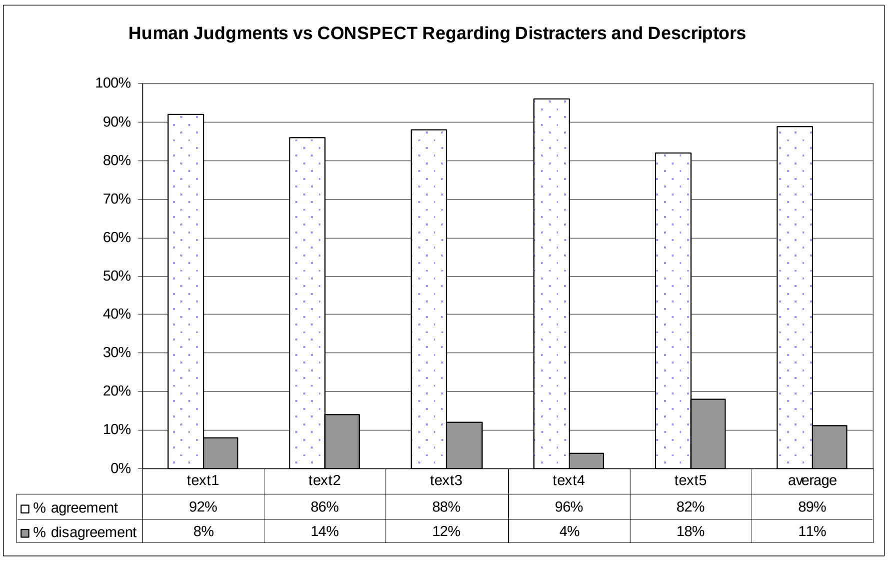

Figure 6. Comparing Human and CONSPECT Judgments

Bibliography

Brandes, U. and T. Erlebach (2005). Network Analysis: Methodological Foundations. Berlin, Heidelberg, Springer-Verlag.

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

Deibel, K. (2005). "UW Card Sort Analyzer: Version 1.1." 2010, from http://www.cs.washington.edu/research/edtech/CardSorts/.

Evans, V. and M. Green (2006). Cognitive Linguistics, An Introduction. Edinburgh, Edinburgh University Press.

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

Goldsmith, T. E., P. J. Johnson, et al. (1991). "Assessing Structural Knowledge." Journal of Educational Psychology **83**(1): 88-96.

Landauer, T. K., D. McNamara, et al., Eds. (2007). Handbook of Latent Semantic Analysis. New Jersey, US, Lawrence Erlbaum Associates.

Randolph, J. J. (2005). Free-marginal multirater kappa: An alternative to Fleiss' fixed-marginal multirater kappa. Joensuu University Learning and Instruction Symposium. Joensuu, Finland.

Randolph, J. J. (2005). "Online Kappa Calculator." from http://justusrandolph.net/kappa/.

Rugg, G. and P. McGeorge (1997). "The sorting techniques: a tutorial paper on card sorts, picture sorts and item sorts." Expert Systems **14**(2): 80-93.

Saeed, J. I. (2009). Semantics, Wiley-Blackwell.

Salton, G., A. Wong, et al. (1975). "A vector space model for automatic indexing." Communications of the ACM **18**(11): 613-620.

Schvaneveldt, R. W., F. T. Durso, et al. (1989). "Network Structures in Proximity Data." The Psychology of Learning and Motivation **25**: 249-284.

Upchurch, L., G. Rugg, et al. (2001). "Using Card Sorts to Elicit Web Page Quality Attributes." IEEE Software **July/August**.