Visualising Textual Knowledge about Risks to Aid Risk Communication

Gary McKeown, Noel Sheehy

School of Psychology
Queen's University Belfast
University Road, Belfast, BT7 1NN
United Kingdom
g.mckeown@qub.ac.uk, n.sheehy@qub.ac.uk

Abstract

This paper demonstrates a potential application for latent semantic analysis and similar techniques in visualising the differences between two levels of knowledge about a risk issue. The HIV/AIDS risk issue will be examined and the semantic clusters of key words in a technical corpora derived from specific literature about HIV/AIDS will be compared with the semantic clusters of those in more general corpora. It is hoped that these comparisons will create a fast and efficient complementary approach to the articulation of mental models of risk issues that could be used to target possible inconsistencies between expert and lay mental models.

1 Introduction

Differences in expertise between professional and lay audiences is a significant barrier to effective risk communication. Similar terms (e.g. risk) may have different meanings for professional and lay groups, leading to potential confusion and the possibility that a message may be interpreted in unintended ways. From the perspective of the expert the process of communicating risks involves deciding on the appropriate message content for the intended audience so that recipients can make an informed judgement about the risks under consideration. In these communications the knowledge of an expert communicator differs substantially from that of a non-expert audience. The difference is partly due to different levels of understanding and different usage of the language contained in the message. Often words used by experts are employed in a technical manner (e.g. hazard), however when such a communication passes to a non-expert audience the same words are understood in a different manner leading to mixed messages, confusion and a failure to communicate the desired message (Jardine & Hrudey, 1997)

A popular approach to overcome these difficulties focuses on articulating the mental models of experts and non-experts and mapping the differences between them (Morgan, Fischhoff, Bostrom & Atman, 2001). Typically a detailed expert influence diagram is created from the technical literature and reviewed by technical experts, then a series of open ended and structured interviews is conducted with the goal of creating a model of a non-expert viewpoint of the same risk. These models are then compared and contrasted in order to highlight the areas in which the differences in knowledge and prospects for misunderstanding are greatest. Risk communications are then targeted at these areas to minimise misunderstandings.

A major limitation of the mental models approach lies in the large investment of labour and time required to produce the expert and lay models. Morgan et al. (2002) argue that this is a small price to pay given the potential enormity of costs if a risk communication is badly formed and delivers the wrong message. A complementary approach is suggested which could lessen the burden associated with the articulation of mental models. Using automatic knowledge extraction techniques a statistical representation of the semantic knowledge in a corpora of information from the expert domain and from a more general literature corpora allows the creation of dual semantic spaces. This can be used by a risk communicator to gauge some of the differences between the technical and a general viewpoint of the same word.

2 Latent Semantic Analysis

Latent Semantic Analysis has been used in a variety of circumstances to elucidate the knowledge contained in a body of textual documents and it is proposed that it provides a statistical representation of that knowledge (Kintsch, 1998; Landauer & Dumais, 1997). Initially used in information retrieval applications it was later adapted for psycholinguistic analysis (Landauer, Foltz & Laham, 1998, Landauer, 1998).). It is an associative technique that considers word co-occurrences and not word order, syntax or rhetorical structure. LSA transforms a document term matrix of word co-occurrences into a high-dimensional (c.200-300) semantic space through singular value decomposition. Thus, the meaning of a word is represented (as a word vector) in a semantic space of approximately 300-dimensions. Additionally sentences, phrases, paragraphs and documents can all be represented in this same space. LSA allows similarity comparisons (measured by the cosine between two vectors) between words, sentences and documents.

Clusters of words can be derived from LSA and similar techniques (Widdows, Cederberg, & Dorow, 2002) which can provide a representation of the surrounding semantic space in which a word exists. To the extent that the corpora represent information from a technical domain and a general domain these clusters can be used to provide lists of semantic neighbours and semantic graphs which can help distinguish the differences between technical and lay conceptions of key words in specific risk issues.

This paper demonstrates a potential application for LSA and similar techniques in visualising the differences between two levels of knowledge about a risk issue. Using the specific risk issue of HIV/AIDS the semantic clusters of key words in the technical corpora will be compared with the semantic clusters of those in the more general corpora.

An evaluation of these comparisons will be used to assess the degree to which this approach can complement traditional techniques for articulating mental models and thereby afford new opportunities to identify potential inconsistencies between an expert and layperson's mental models.

3 Methodology

Morgan et al. (2001) describe an expert model of HIV/AIDS knowledge. Their model is designed predict the factors most relevant to transmission of HIV/AIDS to uninfected people, the health consequences, and the feedback processes involved in further infection. In this paper we are concerned with using latent semantic analysis to produce a semantic network concerning the transmission aspect of this expert model. Figure 1 is adapted from the original model, an influence diagram in which the value at one node is derived from the values of the nodes at the tail of the arrows connected to it (in Figure 1 the original influence lines are faint).

Subjecting a corpus of textual knowledge on HIV/AIDS to a Latent Semantic Analysis provides us with a high dimensional representation of the distance between words and concepts in the body of textual knowledge, also known as a semantic space. It is hypothesised that a corpus derived from literature solely devoted to the area in the expert domain (HIV/AIDS) should contain clusters of words that are closely aligned with the expert influence diagram. Additionally when the same concepts are examined using a semantic space derived from more general literature it is expected that the clusters will bear little resemblance to those in the expert influence model. These differences between usages of words should provide useful information to any communicator of a potential risk concerning the possible areas of confusion between the technical and lay understanding of a concept.

Taking the central concept “transmission” from the expert influence diagram and analysing it using Latent Semantic Analysis provides a list of the nearest neighbours in a high dimensional semantic space (300 dimensions), a separate semantic space exists for each corpus of text. Further, using the techniques of visualising a semantic space suggested by Widdows, Cederberg, and Dorow, (2002) it is possible to provide the communicator with an idea of the semantic clusters involved in both an expert and lay domain of knowledge and perhaps illuminate the differences in a way which is easily and efficiently understood by a communicator. This technique involves using the first order and second order neighbours in a semantic cluster to generate a graph style visual representation of the semantic clustering around a given word or concept. The word transmission will be analysed using two similar techniques, latent semantic analysis and the infomap-nlp program (http://infomap.stanford.edu).

3.1 The Corpora

A number of different corpora are used in this study:
An expert corpus, derived from a number of publications on HIV/AIDS primarily from the US Centre for Disease Control and Prevention, the UCSF Center for HIV information and the Joint United Nations Programme On HIV/AIDS (UNAIDS). These comprise a corpus of 3000 documents with approximately 800,000 words and 20,000 unique words. General reading corpus is an in-house corpus consisting of a large number of news items taken from the Reuters-21578 text categorization collection, and a number of written texts from novels and religious writings. These comprise a corpus of 10,000 documents with nearly 3 million words and 48,000 unique words. The British National Corpus (BNC) as used by infomap-nlp.

4 Results

Table 1 shows the rankings for the neighbouring words of the word ‘transmission’ in various corpora.

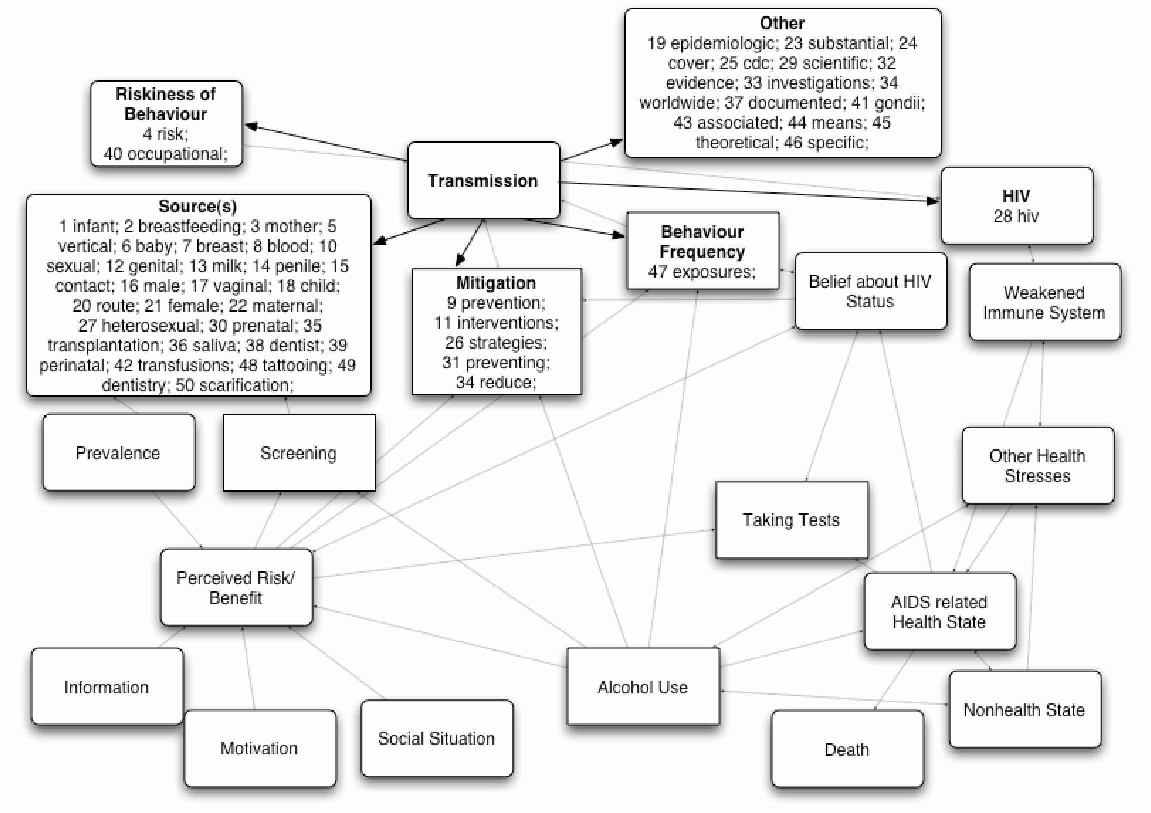
Table 1. Rankings of the words neighbouring the word ‘transmission’ in four corpora.

Rank	*LSA – Expert*	*LSA–General*	*Infomap-NLP – Expert*	*Infomap-NLP – BNC*
1	infant	designs	vertical	reception
2	breastfeeding	Philips	steam	distribution
3	mother	specializing	leroy	engine
4	risk	Varian	chorioamnionitis	generation
5	vertical	Pye TVT	quantitate	processing
6	baby	equipment	potable	creation
7	breast	frequencies	double-gloving	devolution
8	blood	visual	conception	telex
9	prevention	measuring	plummer	behaviour
10	sexual	lcds	centrality	content
11	interventions	Klugt	undercooked	memory
12	genital	Tekronix	replacements	possession
13	milk	displays	nonbreastfed	publication
14	penile	GEC	healthiest	storage
15	contact	Cemax	child	telephone
16	male	Emory	prevention	transfer
17	vaginal	UCLA	milk	transformation
18	child	Pixar	depicted	recording
19	epidemiologic	clinic	vapor	control
20	route	electronics	wellcome	management

There are clear differences in the way in which words are used in the expert corpus and in a more general corpus. Words used in the expert corpus are unambiguously associated with the disease based sense of the word. The senses in which they are used in a more general corpus reflect transmission in a broadcast, business and electronics sense in the general corpus and an automotive and information technology sense in the British National Corpus. Only the ninth neighbour in the infomap-nlp BNC corpus, ‘behaviour’, could be considered to have a link to the disease sense of the word, and the nineteenth neighbour ‘clinic’ in the general corpus, however further investigation reveals this is more due to health and pharmaceutical business connections of some of the corporations in the semantic cluster.

The LSA has proved more useful for some words than for others. Words such as ‘transmission’ are singular with clearly defined meanings in the sense in which they are used in the expert corpus. 36 of the nearest 50 words in the word cluster surrounding ‘transmission’ could be accounted for by the expert model. Figure 1 shows the nearest textual neighbours to the word ‘transmission’.

Figure 1. The expert model with the expert corpus nearest neighbours for the word ‘transmission’ (adapted from Fischoff & Downs, 1997).

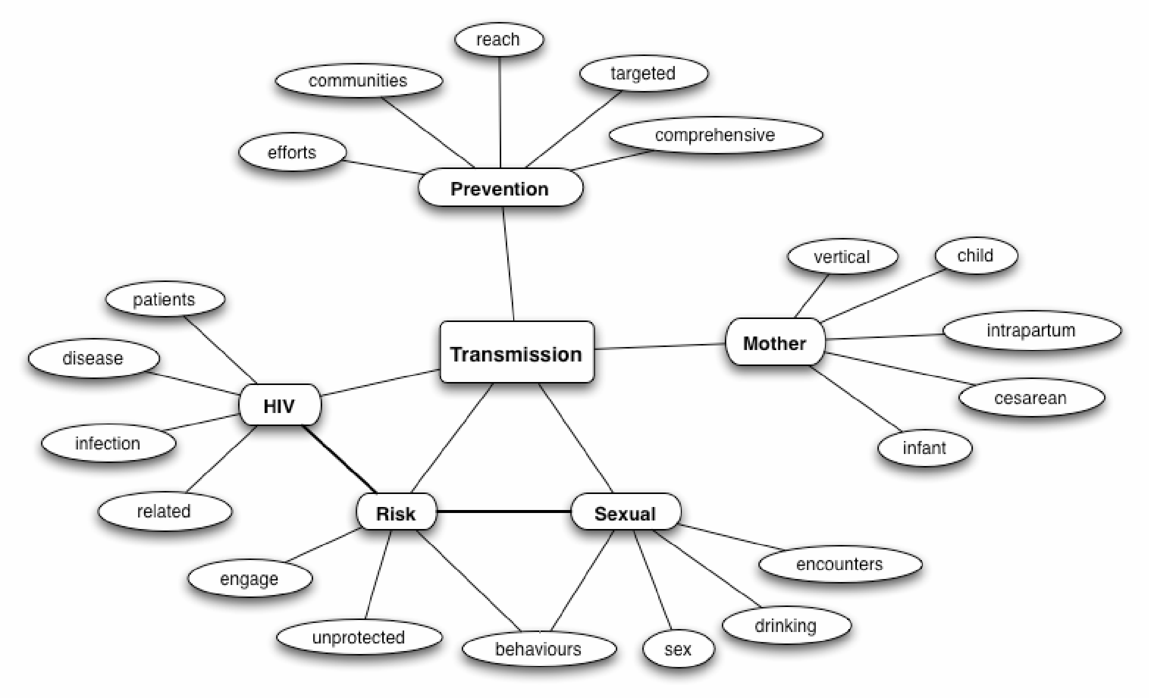

Other
19 epidemiologic; 23 substantial; 24 cover; 25 cdc; 29 scientific; 32 evidence; 33 investigations; 34 worldwide; 37 documented; 41 gondii; 43 associated; 44 means; 45 theoretical; 46 specific;
Riskiness of Behaviour
4 risk;
40 occupational;
Transmission
HIV
28 hiv
Source(s)
1 infant; 2 breastfeeding; 3 mother; 5 vertical; 6 baby; 7 breast; 8 blood; 10 sexual; 12 genital; 13 milk; 14 penile; 15 contact; 16 male; 17 vaginal; 18 child; 20 route; 21 female; 22 maternal; 27 heterosexual; 30 prenatal; 35 transplantation; 36 saliva; 38 dentist; 39 perinatal; 42 transfusions; 48 tattooing; 49 dentistry; 50 scarification;
Behaviour Frequency
47 exposures;
Mitigation
9 prevention;
11 interventions;
26 strategies;
31 preventing;
34 reduce;
Belief about HIV Status
Weakened Immune System
Prevalence
Screening
Other Health Stresses
Taking Tests
Perceived Risk/ Benefit
AIDS related Health State
Information
Alcohol Use
Nonhealth State
Motivation
Social Situation
Death

Less successful are those words which do not have a clear and precise meaning, and those concepts which are made up of more than one word. For instance “Other Health Stresses” assumes prior knowledge on the part of the reader and therefore cannot be easily placed in a semantic space. This is a difficult type of concept for latent semantic analysis, but this difficulty serves to highlight that there are assumptions underpinning this concept and any potential communicator should be aware that this is a source for potential confusion. Words such as ‘Mitigation’ are also difficult. This word is seldom used by itself in the general corpus (normally with respect to prevention and reduction programs in Africa) however in an technical sense it is a useful umbrella term for a number of concepts concerning the reduction, prevention and alleviation of problems associated with HIV/AIDs.

5 Visualization

The nearest neighbours of the semantic model can be used to provide the basis of a map or graph of the semantic cluster in which the word belongs. In creating the graph from the expert corpus a selection was made from the first fifty neighbours as some were clearly referring to a similar concept so the five neighbouring concepts were chosen from two sources, maternal (mother), and sexual, the next two concepts are risk and prevention, also included was the word HIV. A second level of the graph is constructed by using LSA to find the five nearest neighbours for each of transmission's five neighbouring concepts in the semantic space. Figure 2 shows the final graph.

Figure 2. Semantic Graph for 'transmission' derived from the expert corpus.

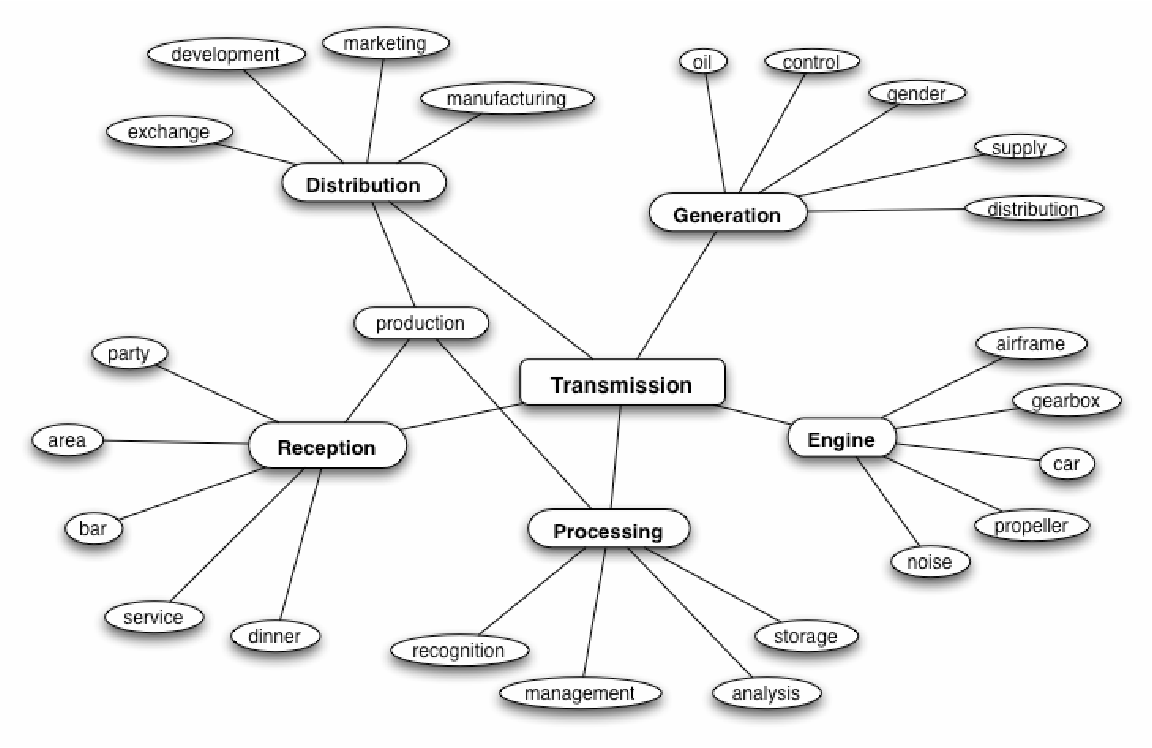


There appear to be three major clusters. One combines elements of HIV, Risk and Sexual. A second concerns Prevention and a third concerns maternal issues. There are some overlaps with the original expert model: the axis of sexual risk and HIV is clearly linked to behaviours and to alcohol (drinking). The mitigation node is covered best by the link to prevention and its neighbours . Morgan et al (2001) say a strategic decision was made when creating their influence model to collapse the main modes of transmission, this is not possible using LSA and is reflected in the prevalence of the modes of transmission as neighbouring concepts. Transfusions risks are absent but would have been present if the next neighbouring concept, blood, had been added in place of HIV. Also absent and of note is the risk of transmission through injecting drug users which features in the corpus. This mode of transmission is not closely associated with the word transmission in the semantic space.

Figure 3 shows a semantic graph for the word 'transmission' derived using infomap-NLP from general literature in the British National Corpus.

The semantic graph in Figure 3 displays a markedly different usage of the word 'transmission': one closer to an engineering and information technology sense of the word. This connotation of diffusion or dissemination may be the first that comes to mind in a lay audience. The discrepancy between these two usages of the word illustrates how latent semantic analysis can identify a potential for poor risk communication.

Figure 3. Semantic Graph for ‘transmission’ derived using infomap-NLP for the BNC.



6 Discussion

The results of the approach seems to be quite promising and the degree of overlap between the LSA derived expert model and the original expert model suggest that further refinement of the technique may be worthwhile. A wholly automatic system would have to take into account the large number of neighbours created for a concept and create fewer and more general concepts. This shows that the skills in creating an expert model lie not only in knowing what is relevant but correctly collating and categorizing groups of concepts into higher order concepts.

The automatic nature of the processing frees the analysis from any bias in deriving the model from the corpus. Bias may of course enter when choosing the items that are used in an expert corpus and care needs to be taken that corpora are balanced, especially expert corpora as they will normally be smaller than more general corpora. If there is confidence that there is a balance of materials in the corpus then this can be used to the advantage of the communicator and can provide an unbiased judge of the relative importance of each issue. There is also value in increasing the confidence in an existing model by comparison with a model derived from an automatic exhaustive search of a literature. In this case however transmission by injecting drug users was not found so the automatic route may not reveal every possibility. Highlighting the various areas may also be useful in providing some starting points for the generation of an expert model.

The utility of comparing a semantic map derived from an expert model with one derived from a more general model depends on the particular nature of a communication, however as a starting point or basic first sketch of the differences between expert and lay conceptions it may have some value.

Further refinements could expand the models by using not only the distances between the various words and concepts which allow the formation of semantic graphs, but also using the measure of the length of a word vector the relevant importance of a concept can be gauged in respect to other concepts in a corpus. Additionally visualization using planar projection another technique used by Widdows, Cederberg, and Dorow (2002) to gauge the difference between usages of words in a bilingual corpus may be useful however this requires parallel corpora with a high degree of overlap and the differences may be too great to be of value in determining the differences between lay and technical usage of a word.

References

Jardine, C.G. & S. E. Hrudey (1997). Mixed messages in risk communication. *Risk Analysis*. 17, 489-498

Fischoff, B., & Downs, J. (1997). “Accentuate the Relevant,” *Psychological Science,* 8(3):1-5

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, MA: Cambridge University Press.

Landauer, T. K. (1998). Learning and representing verbal meaning: The latent semantic analysis theory. *Current Directions in Psychological Science*, 7, 161-164.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato’s problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review,* 104, 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, 25, 259-284.

Morgan, M.G., Fischhoff, B., Bostrom, A. & Atman, C.J., (2001). *Risk Communication: A Mental Models Approach*, Cambridge University Press.

Widdows, D., Cederberg, S., & Dorow, B. (2002). Visualisation Techniques for Analysing Meaning. *Fifth International Conference on Text, Speech and Dialogue,* Brno, Czech Republic, September 2002, pages 107-115