Krippendorff. K. (2004). *Content Analysis. An introduction to its methodology* (2nd ed). Thousand Oaks/London/New Delhi: Sage.
Riffe, D., Lacy, S. & Fico, F.G. (2005). *Analyzing Media messages. Using Quantitative Content Analysis in Research*. Mahwah/London: Lawrence Erlbaum.
Selm, M. van & Hijmans, E. (2006). Digitale documenten. In F. Wester (Red.). *Inhoudsanalyse: theorie en praktijk* (pp. 207-226). Alphen aan den Rijn: Kluwer.

A thematic or a relational approach to the financial crisis?

Commentary to Atkisson, Monaghan and Brent

Jan Kleinnijenhuis

Veyor® is a trademark of Idea Works, Inc. It is a text analysis program that performs, either by itself or in combination with programs such as Qualrus® and Globalpoint®, not only word category counts, but also sentiment analysis. According to a newspaper article about a recent application to a campaign for the US Senate elections (Reed, 2010), the sentiment towards the candidates in blogs and newspapers as extracted by Globalpoint® predicted the outcome of the elections more accurately than a telephone survey. Candidates received positive or negative points based on what was being said about key issues in the race and were categorized under headings such as 'government,' 'economy,' 'personal' and subsets such as 'free market' and 'tax issues'.
Two questions will be addressed in this short review. First, how is Veyor® embedded in the scientific literature? Next, what is the performance of Veyor® in analyzing the KWALON dataset on the economic downturn (Atkisson, Monaghan & Brent, 2010)?

Veyor and the research literature

Atkisson et al. compare Veyor® to existing methodologies in content analysis, text mining and qualitative text analysis (QDA), although I think there is more to say about the three methodologies. For instance, Krippendorff (2004) treats many more approaches to content analysis than the early forms of automated content analysis 'which often uses simple word frequency and keyword-in-context statistics to elucidate the data' and 'relies little on advanced computational techniques'. Already the very first book on automated content analysis had a chapter about the valuation of positive and negative relationships between nations according to the press (Stone, Dunphy, Smith & Ogilvie, 1966). In the section on text mining, I would have welcomed some information on which recent advances in Semantic Web approaches and Natural Language Processing, for example with regard to machine learning (named entity recognition, part-of-speech-tagging, or grammar par-

sing), were helpful to build Veyor®. The aim of QDA is to retrieve meanings from texts with an open mind by letting the texts speak for themselves in an inductive fashion. In practice, QDA researchers sometimes adopt a fairly naïve hierarchical classification approach, by assigning codes to text segments, that could be organized in an hierarchical fashion. A classification allows for answering questions about the *themes* that dominate the discourse (e.g. specific actors, specific issues), but a classification does not help much to answer questions about the *relationships* between them. Given their research questions, qualitative researchers should more often apply a *relational* approach (or semantic network approach) to content analysis (Popping, 2000; Roberts, 1997; Van Atteveldt, 2008) in stead of a *thematic* classification approach. To answer questions about relationships in texts one has to resort to other structures in language than semantic classification trees, e.g. to conjugations and grammar trees, and especially to Subject / Predicate / Object – templates (Van Atteveldt, 2008).
One of the key features of Veyor is its reliance on human coders to classify a sample of sentences, whereupon these classification codes are used as training materials for a *supervised* machine learning algorithm (non-supervised learning algorithms are strictly rule-based and do not rely on human codings).

The description of the supervised machine learning procedure leaves many standard questions from the field of Natural Language Processing unanswered:

1. Which types of textual data (manifest word forms only? manifest classification codes of coders only? word lemma's? part-of-speech-tags? grammatical functions?) are generated in the preprocessing stage before the actual machine classification?
2. Which machine classification algorithm was used (e.g. Naïve Bayes, Support Vector Machines, Maximum Entropy)?
3. Why is the ultimate machine classifier based on the complete set of human codings? It's a common practice to use only a part of the coding data as training materials, to see whether the algorithm can predict the remaining coding data (often this is done in a bootstrap fashion).
4. Apparently Veyor accepts the outcomes of one specific machine classification algorithm as the ultimate classification, thereby neglecting that even these outcomes depend already on the precise training materials, the bootstrapping parameters and the precise settings of a number of highly technical parameters. Accepting only outcomes agreed upon by a variety of machine learning algorithms and hand-coding the sentences on which the algorithms disagreed wildly will presumably lead to more valid research outcomes and to more sophisticated training materials for the algorithms (Hillard, Purpura & Wilkerson, 2008).

Due to the lack of information about the precise computational techniques that were included in Veyor, it's impossible to evaluate the 'engine' of Veyor, but it is still possible to evaluate the program on the basis of its results.
The outcome that a content analysis of social media and newspapers predicts the outcome of an election with two candidates fairly close (Reed, 2010), was illustrated before by

Fan and Tims (1989), who showed that the frequencies of positive and negative words used in newspapers alone sufficed already to predict the outcomes of elections fairly closely. Techniques to extract textual content from the web and social media are well-established in the meanwhile, but the problems which still remain are not mentioned by Atkisson et al., e.g. neither the problem how to distinguish differences in content between social media and newspapers from differences in language style, nor the problem how sampling weights should be attached to specific postings on the web.

Veyor and the financial crisis

The research question for the KWALON study asks: 'what are the primary actors, causes and consequences mentioned in relation to the economic crisis' (Atkisson, et al., 2010). This is a question about *frames*. According to Entman, a frame is a particular emphasis in texts 'to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation' (Entman, 1993). In other words, a frame is a network template that specifies that any given 'problem', e.g. the financial crisis, usually will have a number of causes, a number of treatments (which may or may not relate to the causes), and a number of actors who caused the (causes of) the problem and/or recommended a treatment. Framing assumes a *relational approach*, or *semantic network analysis approach* to content analysis (van Atteveldt, 2008), since we do not know in advance whether a specific theme, such as low interest rates is framed by specific observers as a direct or indirect cause of the financial crisis (e.g. low interest rates » higher debts » lack of economic trust » decreasing stock prices), as a direct or indirect treatment of the crisis (e.g. lowering interest rates » higher investments » economic trust » increase in stock prices), or as a consequence of the crisis (e.g. financial crisis » lack of economic trust » decreasing demand for money/investments » low interest rates) (Kleinnijenhuis, 2008).

Atkisson et al. (2010), however, implicitly apply a thematic approach to content analysis, since they seem to consider each specific theme as either a cause or a consequence by definition. Although 'interest rates' may occur in a variety of different causal chains, their thematic approach prompts them nevertheless to consider 'interest rates' by definition as a *cause* of the crisis. Their classification of causes does not clarify whether 'interest rates' refer to 'high interest rates', 'low interest rates' or to both, although different causal chains assume opposed effects of high and low interest rates. Atkisson et.al. propose a distinction between 'positive consequences' and 'negative consequences', but the above-mentioned causal chains indicate that it Is not a matter of a priori thematic classification, but a matter of framing whether 'low interest rates' would belong to the positive or to the negative consequences.

References

Atkisson, C., Monaghan, C. & Brent, E. (2010). Using computational techniques to fill the gap between qualitative data analysis and text analytics. *KWALON 45*, 15(3), 6-19.

Atteveldt, W. van (2008). *Semantic Network Analysis: techniques for extracting, representing and querying media content*. Charleston, SC: BookSurge Publishers.

Entman, R.M. (1993). Framing: Toward Clarification of a Fractured Paradigm. *The Journal of Communication*, 43(4), 51-58.
Hillard, D., Purpura, S. & Wilkerson, J. (2008). Computer-Assisted Topic Classification for Mixed-Methods Social Science Research. *Journal of Information Technology & Politics*, 4(4), 31-46.
Kleinnijenhuis, J. (2008). Reasoning in economic discourse: a network approach to the Dutch Press. In K. Krippendorff & M.A. Bock (Eds.), *The Content Analysis Reader* (pp. 430-442). Thousand Oaks: Sage.
Krippendorff, K. (2004). *Content Analysis*. Thousand Oaks: Sage.
Popping, R. (2000). *Computer-assisted text analysis*. London: Sage.
Reed, D. (2010). Business profile Idea Works: making sense of online chatter. *Columbia Business Times, July 23, http://www.columbiabusinesstimes.com/8490/2010/07/23/*.
Roberts, C. W. (1997). *Text analysis for the social sciences: methods for drawing statistical inferences from texts and transcripts*. Mahwah: Erlbaum.
Stone, P.J., Dunphy, D.C., Smith, M.S. & Ogilvie, D.M. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. Boston: The MIT Press.

Using Computational Techniques to Fill the Gap between Qualitative Data Analysis and Text Analytics

Response to commentary

Curtis Atkisson, Colin Monaghan and Edward Brent

The authors would like to thank both Martine van Selm and Jan Kleinnijenhuis for their thoughtful comments regarding our paper. They each raise a number of interesting points. The comments focus broadly on three areas, and we will structure our response around those three areas. We will first address comments on the literature review, then the Veyor system and finally, comments on the analysis itself.

The literature review

As Van Selm mentions, detailing strengths and weaknesses of different methodologies is a useful tool for understanding when a certain methodology is well suited to address a specific research question or class of research questions. Our review of strengths and weaknesses is not intended to be a comprehensive listing of possible research questions and their relative strengths, but rather to be a basic guide to understanding the differences between three techniques. The point Van Selm makes regarding evaluating the validity of a research method in context of the research question is a good point, but that does not address the statistical validity of the Veyor system, which we evaluate in the paper.