Combining human coding and automated coding in Veyor®

Commentary to Atkisson, Monaghan and Brent

Martine van Selm

Atkisson, Monaghan and Brent review in 'Using Computational Techniques to Fill the Gap between Qualitative Data Analysis an Text Analytics' three methods (qualitative data analysis, content analysis and text mining) that are used in order to examine streams of digital textual materials that are increasingly accessible through the Internet.

The authors structure their methods review by describing strengths and weaknesses of each method. Weaknesses and strengths center around issues of efficiency (speed and effort), freedom left to the researcher to develop new ideas from the empirical materials ('coding up'), and the general criteria for research quality: validity and reliability.

Mapping out strengths and weaknesses of a method phenomena can be insightful, and the authors' description is so. However, when it comes to research methods in general, their strengths and weaknesses are probably best considered when linked to a specific research question. Inasmuch as a research question determines what research method is appropriate, it also is decisive for a method's performance on validity and reliability. For example open interviewing is valid in a study on the *meaning of media use*, whereas observation or a diary method are so in a study on *actual media behaviors*.

Besides this point of criticism, the authors do not bother themselves with research aspects that are specific for the analysis of web materials. Examples are archiving material from dynamic data sources such as the Internet, or defining workable recording units (see Van Selm, 2006).

In the second part of their contribution, Atkisson and colleagues introduce the software package Veyor in a brief manner and emphasize its ability of combining human coding with automated coding. The authors emphasize that in Veyor human researchers and the computer are assigned to tasks for which they are best. This combination is not new but has been applied in for instance the NET-method (Van Cuilenburg, Kleinnijenhuis & De Ridder, 1989). Atkisson and colleagues emphasize the suitability of this package for the analysis of streams of web materials. More in particular they describe how the software performs in a specific research task. The task involves examining world wide views of the economic collapse as revealed in traditional and social media sources. The researchers used 97 newspaper articles (traditional medium) and 102 blog entries (social medium) adding up to more than 5800 blogged sentences. The research question addressed (What are the primary actors, causes and consequences mentioned in relation to the economic crises?) actually does not invite much to a qualitative analysis, as it is not a question about the meaning of the crises. Instead, the question could

be addressed by counting the instances in which various types of actors, causes and consequences can be identified in the sample of newspaper articles and blogs. This is also basically what Atkisson and colleagues do in their analysis, as will be shown below. Such an approach requires however a well conceived sample that is able to represent the phenomenon under study (newspaper articles and blogs about the economic crisis in a specific period) in one way or another. Careful procedures for sampling media materials have been described extensively in communication science literature (see for instance Riffe, Lacy & Fico, 2005).

Atkisson and colleagues subsequently explain how they used the computer in their analysis. In general the use of computers in text- and content analyses has gained great importance. Computers are capable of dealing with large amounts of textual materials, and of working in a fast and rigid manner. At the same time, computers lack the capability of assigning meaning to texts as they are no competent language users, something humans naturally are. Instead, computers are (only) capable of applying pre-defined rules. Computers have been used in text- and content analysis in two ways, roughly spoken (see for instance Krippendorff, 2004; Riffe et al., 2005). The first way involves the use of computers as text processor in which their main task is counting words or highlighting key words (e.g., in Key Words In Context-procedures). The second way involves computerized analyses of texts, in which the computer is programmed in order to generate conclusions about the content of a text. Atkisson and colleagues' approach in Veyor is an example of a computerized text analysis, and more specific of a thesaurus/dictionary approach. During the phase of human (open) coding the researchers created a study-specific dictionary or thesaurus (a code scheme containing words/concepts and rules describing how to apply these), by which the computer decides whether a recording unit (sentence) fits into one of the three predefined categories of actors, causes and consequences. The analyses do not go beyond a thematic analysis of the frequencies of various actors, causes and consequences (see Kleijnnijenhuis & Van Atteveldt, 2006). Especially regarding the latter theme the results of the analysis could have been more informative: what makes a consequence of the economic crisis negative, neutral of positive? In addition, it would have been interesting to see how Veyor supports the analysis of correspondence between for instance type of actor and cause, or between type of cause and consequence. This is even more interesting because Veyor can automatically generate cross tabulations. Inasmuch as the aim of this article was to describe Veyor's suitability for the analysis of streams of web materials, an analysis of methodological differences between analyzing traditional newspaper articles in Veyor versus the Internet blogs would have been a welcome addition.

References

Cuilenburg, J.J. van, Kleinnijenhuis, J. & Ridder, J.A. de (1989). *Tekst en betoog. Naar een gecomputeriseerde inhoudsanalyse van betogende teksten*. Muiderberg: Coutinho.

Kleinnijenhuis, J. & Attenveldt, W. van (2006). Geautomatiseerde inhoudsanalyse, met de berichtgeving over het EU-referendum als voorbeeld. In F. Wester (Red.), *Inhoudsanalyse: theorie en praktijk* (pp. 227-250). Alphen aan den Rijn: Kluwer.

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-