Choice of Text Analysis Software in Organization Research: Insight from a Multi-dimensional Scaling (MDS) Analysis

Vincent J. Duriau[1], Rhonda K. Reger[2]

[1,2]ITAM, Av. Camino a Santa Teresa, #930, Del. Magdalena Contreras, México, D. F. 10700
vduriau@itam.mx, vduriau@rhsmith.umd.edu
[2]R.H. Smith School of Business, University of Maryland, College Park, MD 20742-1815
rreger@rhsmith.umd.edu

Abstract

Content analytic approaches have made great strides in organization studies in the course of the past two decades. However, management scholars are often challenged to choose the best software solution to implement their research project. In order to assist in such a critical methodological decision, we develop a more thorough and comprehensive categorization of computer-aided text analysis (CATA) software.

First we summarize the strengths and limitations of CATA as it has been applied in organization studies. Second, we review typologies of such software that have been proposed. Finally, we report on the results of a multidimensional scaling (MDS) analysis of 33 CATA packages. These findings should help management researchers to make better informed decisions in their choice of CATA software.

Résumé

Les approches d'analyse de contenu ont fait d'importants progrès pour la recherche en gestion au cours des deux dernières décennies. Néanmoins, les chercheurs en management ont souvent des difficultés à choisir le meilleur logiciel pour leur projet. À cet effet, nous développons une catégorisation plus rigoureuse et complète des logiciels d'analyse de contenu.

Nous commençons par résumer les avantages et les limitations des logiciels d'analyse de contenu dans leur application à la recherche en gestion. Ensuite, nous résumons les typologies existantes de ces logiciels. Finalement, nous décrivons les résultats d'une analyse « multidimensional scaling » de 33 logiciels d'analyse de contenu. Ces résultats devraient assister les chercheurs en gestion à prendre de meilleures decisions dans leur choix de logiciels.

Keywords: computer-aided text analysis, multidimensional scaling.

1. Introduction

The computer revolution has contributed to the proliferation of content analytic methodologies in organization research (Weitzman and Miles, 1995; Kelle, 1995; Roberts, 1997; Tesch, 1991). New programs with enhanced capabilities have made subtle analysis of large amounts of quantitative and qualitative data possible (Gephardt, 1993; Lissack, 1998). We review the computer-aided text analysis (CATA) packages available today and use multidimensional scaling (MDS) to make sense of this vast array. With the breathtaking pace of new feature introductions, we use the MDS results to develop criteria that can help management researchers evaluate future software releases.

Our interest centered on two major questions: 1) How have CATA software been applied in organization studies? And , 2) How can CATA packages be categorized to help management researchers make a better informed decision to implement their projects?

2. Computer-aided Text Analysis in Management Research

2.1 *Definition of computer-aided text analysis*

Several definitions of computer-aided text analysis (CATA) have been proposed (see, e.g., Kabanoff, 1997; Mossholder *et al.*, 1995; Wolfe *et al.*, 1993). Since multiple methodologies and technologies have been included under the CATA rubric, we have adopted Wolfe *et al.*'s inclusive definition: CATA is constituted by software programs that "facilitate the analysis of textual data" (Wolfe *et al.*, 1993, 638). Most authors use the expression computer-aided textual analysis (CATA) because its short form is preferred to the unfortunate acronym for computer-aided content analysis. Nonetheless, the terms are interchangeable.

2.2. *Strengths of CATA*

One way to understand the contribution of computers to content analysis is to contrast computer-aided and manual approaches (Kelle, 1995). In addition to easier data manipulation, the use of software affords several analytical advantages that greatly enhance the methodology. First, computerization allows the manipulation of large data sets (Gephart, 1991; Lissack, 1998; Morris, 1994; Wolfe *et al.*, 1993). The complexity and potential interrelationships of concepts increase exponentially with the quantity of data. Software programs offer features for organizing, searching, retrieving, and linking data that renders the process of handling a large project much more manageable. For instance, Lissack (1998) described how parsing software can be used to sample concepts from a large corpus. This sampling approach allows the researcher to content analyze a reasonable amount of data representative of the initial corpus.

Second, computers reduce the time and cost of undertaking content analytic projects (Mossholder *et al.*, 1995). Time savings stem from the minimization of the coding task, the reduction in coder training, the elimination of inter-rater checks, and the ease of running multiple analyses (Carley, 1997).

Third, the use of computers addresses several of the reliability concerns associated with manual coding (Morris, 1994; Gephart and Wolfe, 1989; Wolfe *et al.*, 1993). Coding rules are made explicit which ensures perfect reliability and comparability of results across texts.

There are encouraging results that the semantic validity possible with manual coding using multiple coders can be achieved at a lower overall cost with CATA (Kabanoff, 1996; Kelle, 1995; Morris, 1994). Morris (1994) tested the validity and reliability of manual and computerized approaches. Using mission statement data from Pearce and David (1987), she compared the outcome of computerized coding in ZyIndex, a text management software, with that of a panel of six graduate business students. She found that results from ZyIndex and the human coders agreed at an acceptable level and that computerized coding yielded an acceptable level of semantic validity (Morris, 1994).

Finally, the use of network concepts have been one of the most exciting developments of the past few years in CATA research (Kelle, 1995). New linkage features between text, memos, and codes such as hyperlinks and graphical tools apply to the areas of theory building, hypothesis testing, and integration of qualitative and quantitative analysis. These developments of CATA seem particularly apt to quell concerns about the decontextualization of results that is inherent to a methodology based on coding and retrieval (Dey, 1995; Prein and

Kelle, 1995). Gephart (1993) also observed the methodological fit of computer-aided text analysis with grounded theorizing because information can be retrieved in meaningful ways that allows for the emergence of grounded hypotheses (see also, Wolfe *et al.*, 1993).

2.3. *Limitations of CATA*

There are still a number of questions associated with the use of CATA (Carley, 1997; Gephardt, 1991; Morris, 1994; Tallerico, 1991). The debate regarding manifest versus latent content comes first and foremost (Woodrum, 1984). Although CATA software is increasing in sophistication, measuring content with computers will miss some latent aspects within texts such as tone or irony of expression (Morris, 1994). However, human coders also exhibit low reliability for latent content. Further, we believe that the significance of these problems may be over-estimated for business texts appearing in corporate documents. These documents are usually written for clarity because they are read by people from around the world. Still, validity may be a particularly critical issue when dealing with metaphors, homonyms, colloquialisms (Carley, 1997), and other aspects of natural language (Morris, 1994).

There are three additional issues in implementing CATA. First, retrieval capabilities may be insufficient in certain software categories (Gephart and Wolfe, 1989). Second, researchers should avoid the pitfalls of the false sense of security afforded by a computer and the justification fallacy of using a computer program (Tallerico, 1991). Computerization will never replace human judgment in all cases. Finally, the fragmentation of the field of CATA software makes the choice of the appropriate package difficult (Kabanoff, 1996).

Categorization of CATA Software

Several interesting and important CATA typologies have been proposed. For instance, Tesch (1991 and 1990) introduced a typology based on the two dimensions of methodology and technology that has been adapted by several other authors (see, e.g., Wolfe *et al.*, 1993; Roberts, 1997). She made the – now blurring – distinction between commercial and academic software (Tesch, 1991 and 1990). Then, she proceeded with categorizing the academic packages.

More recently, Weitzman and Miles (1995) established a practical list of types of CATA packages, ranging from simple to more complex programs. The typologies of Tesch (1991) and Weitzman and Miles (1995) are summarized in Table 1.

Tesch (1991)

Type of software, main features	Examples*
Descriptive/interpretive analysis. The main goal is to discover the meaning of the phenomena (patterns, types). Data is kept unstructured and is flexibly manipulated. Main functions are coding and retrieval of text. Enhancements may include frequency count, chain, co-occurrence, advanced search, memoing, and quantitative option.	Ethnograph, Qualpro, Textbase Alpha, TAP, MARTIN, LTT Ethnoscript
Theory-building research. The objective is to elicit concepts/linkages in the data. Basic functions of text coding and retrieval are included. Key functions are concept identification, co-occurrence.	Kwalitan, HyperRESEARCH, NUDIST, AQUAD, ATLAS.Ti

Traditional content analysis or cultural analysis. Expansion of the traditional search function of word processors.	General Inquirer, TextPack, Word-Cruncher, FlexText
Linguistic Programs. Linguistic aspects of data.	CODEF, PLCA

* In 1991, these programs were running on MS-DOS or MacIntosh personal computers. Examples are given for the MS-DOS platform only.

Weitzmann and Miles (1995)

1. W*ord processors*, such as Word and WordPerfect, can be used for a variety of support tasks including handling field notes, transcribing interviews, preparing files, taking project notes (memoing), and writing reports.
2. *Text retrievers* specialize in finding and retrieving words, phrases, and characters strings. Some, such as ZyIndex, have content analytical features.
3. *Text base managers* are focused on the organization of textual data, including features such as sorting, searching, and retrieving. Asksam and WinMax are examples.
4. *Code and retrieve programs* are geared toward the coding and manipulation of text. Ethnograph is an example.
5. *Code-based theory builders*, such as NUD*IST, provide additional features including theory building tools, such as connections between codes, coding hierarchies, and formulation and testing of hypotheses.
6. *Conceptual network builders* focus on the network aspects of theory development and provide graphical features to support theoretical development

Table 1. CATA Typologies

Finally, Kabanoff (1997) suggested in his introduction to the *Journal of Organizational Behavior*'s special issue a typology of CATA-based management research using two dimensions: data sources and analytic methods. Data sources can be evoked – collected by the researcher in questionnaires or surveys, or natural – documents such as annual reports or newspaper articles. The second dimension ranges from quantitative to qualitative (Kabanoff, 1997 and 1996).

3. MDS Analysis of CATA Packages

Given the challenges facing management scholars to choose the best software solution to implement their content analytic project, we developed a more thorough and comprehensive categorization of existing computer-aided text analysis (CATA) software. We conducted an MDS analysis comparing the CATA packages referenced on the Georgia State University (GSU) web site on content analysis (http://www.gsu.edu/~wwwcom/content/csoftware/software_menu.html), using the dimensions from the three typologies discussed above as well as additional practical considerations.

3.1. *Methods*

It is difficult to provide a comprehensive list of all existing CATA software and the GSU web site was the most exhaustive reference that we could find. Among the 57 CATA packages described on the GSU web site, 24 were excluded from the MDS analysis because they did

not operate on the Windows operating system, were no longer supported, had a functional scope that was too broad, or did not have an English version.

We then established criteria on which to evaluate these packages. The choice of features was based on extant typologies (Kabanoff, 1997; Tesch, 1991; Weitzman and Miles, 1995) as well as a review of the documentation downloaded from CATA software suppliers' web sites. A comprehensive database was compiled using web sites' information, software demo versions, users' manuals, and contacts with software suppliers by e-mail. Due to its size, the complete database cannot be rendered in this paper, but an evaluation of dtSearch, TATOE, and NVivo according to 30 major features is shown on Table 2.

SOFTWARE PACKAGE	DTSEARCH	TATOE	NVIVO
Price	$199	Free	$425
Orders	1-800-483-4637	N/A	Online/phone
Technical support	(703) 413-3670	N/A	805-499-1325
Website	www.dtsearch.com	http://www.darmstadt.gmd.de/~rostek/tatoe.htm	www.qsr-software.com
Fax	(703) 413-3473	N/A	805 499 0871
E-mail	sales@dtsearch.com	alexa@zuma-manheim.de; rostek@darmasdt.gmb.de	nudist@sagepub.com
User group	N/A	N/A	YES
Demo version	YES	YES	YES
Data Input			
ASCII	YES	YES	YES
Productivity software	YES	N/A	N/A
HTML	YES	YES	N/A
Other	N/A	XML	RTF
Text manipulation			
Preparation	YES	YES	YES
Memoing/coding	N/A	YES	YES
Coding	N/A	YES	YES
Dictionaries	YES	YES	YES
Project management	N/A	YES	YES
Functionalities			

Searching	YES	YES	YES
Basic text statistics	YES	YES	YES
Content Analysis	N/A	YES	N/A
Qualitative Data Analysis	N/A	N/A	YES
Statistical Analysis			
Multiple regression	N/A	N/A	N/A
Cluster analysis	N/A	N/A	N/A
Multidimensional scaling	N/A	N/A	N/A
Other	N/A	N/A	N/A
Data output			
ASCII	N/A	N/A	YES
Productivity software	N/A	N/A	YES
HTML	N/A	YES	N/A
Statistical packages	N/A	YES	YES
Other	DBV, CSV, RTF	XML	RTF

Table 2. Comparison of dtSearch, TATOE, and NVivo

To prepare the data for statistical processing, we indicated the presence of each feature in a software package by a binary count. To simplify the MDS analysis, features were then grouped in seven categories: price, technical support, data input, text management, content analytic features, statistical analysis, and data output. Consistent with extant typologies, we maintained the distinction between the functionalities of searching, basic text statistics, content analysis, and qualitative data analysis.

3.2. *Results*

The results of the MDS analysis confirmed and enriched the typologies previously discussed. First, the two-dimension analysis revealed several categorization dimensions: quantitative-qualitative, standard-advanced, unbundled-integrated, and economic-expensive (see Figure 1). Second, three clusters differentiating qualitative, quantitative, and search CATA packages also appeared, which were consistent with Tesch's (1991) and Weitzman and Miles' (1995) typologies. These results countered our expectations that the market for CATA packages was converging and that the segmentation valid a decade ago was blurring. In addition, CATPAC, SphinxSurvey, and TextSmart clustered into a new group of integrated content analysis packages. While the use of three dimensions improved the quality of the MDS analysis (stress=0.158 and R^2=0.901 for three dimensions, versus stress=0.254 and R^2=0.814 for two dimensions), we show the results for the two-dimension model, which provides a simpler taxonomy of CATA software packages for practical purposes.

4. Conclusion

We provide new insight for the evaluation and selection of CATA software. We identified and empirically tested the typologies proposed in the literature. The results of the multidimensional scaling (MDS) analysis largely confirmed the utility of previously proposed typologies, but also suggested important refinements. Additionally, the placement in two-dimensional space of 33 current and widely available CATA packages should benefit researchers in selecting the solution that is optimal for their project.

The advent of CATA software has led to significant benefits in terms of cost, flexibility, and reliability, while maintaining satisfactory levels of validity (Morris, 1994). In our opinion, the higlt quallyllhalus(l)-9idalcoofco
puter automation and human intervention to increase efficiency and perform more subtle and

Gephardt R.P. (1991). Multiple approaches for tracking corporate social performance: Insights from a study of major industrial accidents. *Research in Corporate Social Performance and Policy*, vol. (12): 359-383.

Gephardt R.P. and Wolfe R.A. (1989). Qualitative data analysis: Three micro-supported approaches. In *Academy of Management Proceedings*: 382-386.

Kabanoff B. (1997). Introduction. Computers can read as well as count: Computer-aided text analysis in organizational research. *Journal of Organizational Behavior*, vol. (18): 507-511.

Kabanoff B. (1996). Computers can read as well as count: How computer-aided text analysis can benefit organizational research. *Trends in Organizational Behavior*, vol. (3): 1-21.

Kelle U. (1995). *Computer-aided qualitative data analysis: Theory, methods, and practices*. Sage Publications.

Lissack M.R. (1998). Concept sampling: A new twist for content analysis. *Organizational_Research Methods*, vol. (1): 484-504.

Morris R. (1994). Computerized content analysis in management research: A demonstration of advantages and limitations. *Journal of Management*, vol. (20): 903-931.

Mossholder K.W., Settoon R.P., Harris S.G. and Armenakis A.A. (1995). Measuring emotion in open-ended survey responses: An application of textual data analysis. *Journal of Management*, vol. (21): 335-355.

Pearce J.A. and David F. (1987). Corporate mission statements: The bottom line. *Academy of Management Executive*, vol. (1): 109-116.

Prein G. and Kelle U. (1995). Using linkages and networks for theory building. In Kelle U. (Ed.), *Computer qualitative data analysis: Theory, methods, and practice*: 69-79. Sage Publications.

Roberts C.W. (1997). *Text analysis for the social sciences: Methods for drawing statistical inferences from text and transcript*s. Lawrence Erlbaum Associates.

Tallerico M. (1991). Applications of qualitative analysis software: A view from the field. *Qualitative Sociology*, vol. (14): 275-285.

Tesch R. (1991). Introduction. *Qualitative Sociology*, vol. (14): 225-243.

Tesch R. (1990). *Qualitative research: Analysis types and software tools*. The Falmer Press.

Weitzman E.A. and Miles M.B. (1995). *Computer programs for qualitative data analysis*. Sage Publications: Thousand Oaks.

Wolfe R.A., Gephart R.P. and Johnson T.E. (1993). Computer-facilitated qualitative data analysis: Potential contributions to management research. *Journal of Management*, vol. (19): 637-660.

Woodrum E. (1984). "Mainstreaming" content analysis in the social science: Methodological advantages, obstacles, and solutions. *Social Science Research*, vol. (13): 1-19.