Policy & Internet
www.policyandinternet.org

Volume 3, Issue 4 • 2011 • Article 5

Coping with the Cornucopia: Can Text Mining Help Handle the Data Deluge in Public Policy Analysis?

Aude Bicquelet, *LSE*
Albert Weale, *UCL*

Bicquelet, Aude and Weale, Albert (2011) "Coping with the Cornucopia: Can Text Mining Help Handle the Data Deluge in Public Policy Analysis?," *Policy & Internet*: Vol. 3: Iss. 4, Article 5.
DOI: 10.2202/1944-2866.1096
Available at: http://www.psocommons.org/policyandinternet/vol3/iss4/art5

Coping with the Cornucopia: Can Text Mining Help Handle the Data Deluge in Public Policy Analysis?

Aude Bicquelet, ***LSE***
Albert Weale, ***UCL***

Abstract

Information and communication technologies such as the Internet have created a plethora of opportunities for the participation of citizens in policymaking. In the United Kingdom, for instance, this trend has emerged at national and local levels, in domains as diverse as Education, the Environment and Health Care. Given a general renewal of interest in incorporating public opinion into policymaking, devices such as online consultations and electronic surveys have rendered the appeal to 'the people' *seemingly* easier. But an important problem arising from involving the public in decision-making exercises through large-scale electronic participatory devices is the amount of textual data generated. Although there is now a large body of literature devoted to commentating and analysing ways in which politicians *ought to* be involved in listening and responding to public participation in decision-making, issues pertaining to the implementation of such exercises in the light of the volume of information that they produce have largely remained unexplored. In this paper, we assess the potential benefits and shortcomings of using Text Mining methods for the analysis of large-scale consultations submitted via Internet. To this end, the paper reports on the application of computer-aided text analysis to a public consultation organised by the National Institute for Health and Clinical Excellence (NICE) in 2008 on 'End of Life Medicines'.

KEYWORDS: text mining, e-consultation, health care

Author Notes: This research was supported by the Economic and Social Research Council (PTA-026-27-2431).

Introduction: Public Participation in Decision Making and the Emergence of New Technologies

Policy makers in the twenty-first century must contend with two inescapable phenomena. On the one hand, over the last decade or so, there has been a major shift in the policies and practices of national governments concerning the increased attention to, and use of, citizen engagement strategies as a basis for developing more participatory forms of governance (Reddel and Woolock 2004, 76). In the UK, for instance, the Blair New Labour government popularized a number of reforms centered on the ideas of "inclusion" and "partnerships" in domains such as transport planning and health care that have been acclaimed as great successes by policy makers and politicians alike (Rowe and Frewer 2005, 251). Current interest in more engaged, collaborative, and community-focused public policy is also evident in an international context. For example, the 2001 OECD report *Citizens as Partners* concluded:

> "[…] democratic governments are under pressure to adopt a new approach to policy-making—one which places greater emphasis on citizen involvement both upstream and downstream to decision-making. It requires governments to provide ample opportunity for information, consultation and participation by citizens in developing policy options prior to decision-making and to give reasons for their policy choices once a decision has been taken" (OECD 2001, 71).

On the other hand, a significant proportion of government activities have now moved online. Whether it is to provide voters with information on the next general election, or to advise citizens on how to deal with lost or stolen passports, the Internet has brought about, in Margetts's words: "a change to the whole information environment within which government operates" (Margetts 2009, 6). If the Internet has become the main medium of information, it has also become the main medium of interaction between government and citizens. Numerous websites offering opportunities for online democratic participation have blossomed recently. The UK's Hansard Society, for instance, has regularly run e-consultations on behalf of select committees, for example on the Climate Change Bill (2007), the Human Tissue and Embryo Bill (2007), and on the issue of Parliamentary Representation (2010).[1] UK councils and boroughs also regularly invite citizens to take part in online consultations on issues affecting their area. In August 2010, for instance, after putting the issue to residents through an online

[1] http://forums.parliament.uk/html/index.html.

consultation, the London Borough of Hammersmith and Fulham decided to close all Sex Entertainment Venues.[2]

Attention has also been paid to online participation in academic circles, where a strong advocacy for greater public participation via the Internet has grown. Whether in terms of "Strong Democracy" (Barber 1984), "Teledemocracy" (Toffler and Toffler 1995), or "Electronic Republicanism" (Grossman 1995), contemporary advocates of direct democracy have offered a variety of "e-models" for political participation and decision making (see, for instance, McLean 1989; Budge 1996). Although the approaches differ from one another in terms of how to use new technologies for democratic governance, they all share a common set of assumptions:

a. The capacity to store vast quantities of information and the feasibility of gathering and disseminating information via such means as the Internet and associated devices considerably enhances the potential of the public to become more knowledgeable and informed concerning policy and politics.
b. Information and communication technologies (ICTs) enable stronger networking and engagement between citizens and elected officials, increasing the former's capacity to participate in the political process; a domain hitherto more exclusive to the latter.
c. The hazards of misinformed voters making poor choices or decisions on a particular issue can be reduced by improved deliberations with fellow citizens and/or engagement with political experts or their views.
d. The advantages of electronic participatory democracy include a reduction of time and effort traditionally required to participate at the polling station in person; a democratizing effect of new technologies, for example in overcoming problems of social exclusion, especially for those with limited mobility; and greater clarity concerning decisions made, with better information enabling citizens to match their preferences more closely to the choices available, translating into more accurate decision making.

Yet, if representative democracy has "lost its technical monopoly"—because of the advent of new technologies and the rising level of expertise among citizens (Kellner 2008)—free and fast access to information, and free and fast access to new modes of participation also raise new challenges. More precisely, the confluence of the above two phenomena—growing enthusiasm for citizen participation and growing use of the Internet to enable such participation—poses two interrelated challenges for the design and analysis of public policy.

[2] http://democracy.lbhf.gov.uk/mgConvert2PDF.aspx?ID=6694.

DOI: 10.2202/1944-2866.1096

The first challenge is *responsiveness*. Governments and organizations need to be able to demonstrate that all opinions, as expressed through participatory exercises, have been duly considered and carefully weighted before a decision is reached (Pratchett 1999; Needham 2002). However, politicians and bureaucrats are often thought to adopt initiatives such as consultations and discussions in a tokenistic way, but then to ignore the full range of public opinion in the formulation and implementation of law and policy. Tomkova (2009), for instance, has argued that outcomes of e-consultations have been poorly and arbitrarily integrated in the policies they intended to inform (for a good example of this criticism applied to upstream engagement in healthcare, see Rowe and Shepherd 2002; Harrison, Barnes, and Mort 1997).

Of course, this is not a new issue. Yet, if questions such as how to provide adequate feedback on a participatory exercise, or how best to integrate citizens' views in public policy decisions, are perhaps as old as politics itself, they have become even more pressing with the rise of new ICTs; more information and more opportunities to get involved in decision making have generated more expectations from citizens, while feedback and output from public officials are expected to be fast and accurate (Roberts 2004; Walters, Aydelotte, and Miller 2000).

This relates to a seemingly more mundane, but nonetheless crucial challenge: the need to ensure that adequate structures are in place to deal with increased participation via the Internet and other ICTs. In other words, as well as the demands for greater legitimacy through increased citizen engagement, policy makers also need to deal with the practical aspects of new interactive participatory exercises. This problem has not escaped the attention of analysts and Internet experts (Margetts 2009). Whether it is in the form of e-consultations and e-surveys, or in the form of comments left on social networks such as Facebook, political blogs, or emails, the Internet provides a formidable opportunity to constantly gauge public opinion on policy issues. An example of these growing concerns is the *Economist*'s recent 14-page Special Report on how to handle the "Data Deluge" (February 2010). The report suggested that "information management software", also known as text mining, could usefully assist the analysis of large corpora and datasets. It also pointed out, however, that whereas businesses and industries (such as Total or Hewlett-Packard) have already exploited the advantages of text mining to get to grips with the huge amount of numerical and algebraic data being made available on the Internet, governments and public bodies are still lagging.

This article evaluates the benefits to be gained by both academics and policy makers in using text mining for the analysis of public consultations that take place online, or where other data—for example, hard-copy written submissions—are captured digitally. We employ computer-assisted textual

analysis, more precisely the Alceste software, to analyze the National Institute for Health and Clinical Excellence (NICE)'s 2008 public consultation on end-of-life medicines. Our aim is neither to assess the performance of the Alceste software specifically,[3] nor to draw comparisons between different text mining methods. In an earlier phase of our research we made comparisons between automated and semi-automated text mining techniques and underscored their value for the analysis of large corpora.[4] In light of this comparison, our aim here is to provide an illustration of the benefits to be gained from such methods and to reflect on their drawbacks for the analysis of e-consultations. The next section briefly describes the background of the consultation. We then describe our method of analysis and present the principal results, and finally offer an assessment of text mining techniques for the analysis of e-consultations.

Background to the Consultation

An important area where the UK Government has variously intervened to require or encourage greater public involvement is health care. As early as the 1980s, the Griffiths Report and the White Paper *Working for Patients* (DH 1989) championed public participation in planning and priority setting. The use of various devices for public involvement, such as opinion polls, surveys, and focus groups, was recommended in the 1992 White Paper *Local Voices*, while the NHS and Community Care Act of 1990 required local authorities to consult with the public over community care plans (NHSME 1992; NHSCCA 1990). Since then, the patient–professional relationship has been further defined in terms of a partnership, with a gradual increase of patient input in healthcare decision making (Rowe and Shepherd 2002, 275-276; Barnes 1999; Milewa, Valentine, and Calnan 1998).

The UK's National Institute for Health and Clinical Excellence (NICE) was established in 1999. Its role is to make recommendations to the National Health Service (NHS) and purchasing authorities on new and existing medicines, and on treatments for specific diseases and conditions. It is also in charge of developing the standards for appraising healthcare technologies, and in particular whether they are cost effective. From its inception, NICE has involved all its stakeholders (patients, carers, the public, health professionals, and industry) in its activities. Draft scopes of individual guidance and the draft guidance itself are subject to public consultation.

[3] See Bicquelet (2009) for an assessment of the Alceste software for the analysis of large corpora.
[4] See Bara, Weale, and Bicquelet (2007).

DOI: 10.2202/1944-2866.1096

When public consultation begins, NICE publishes the consultation document on its website for four weeks. At the same time, all parties that have registered an interest are informed by email that a consultation has begun. During this process, anyone may submit comments via NICE's website, by email, fax, or post. This is perhaps the most basic and widespread form of e-consultation as per Tomkova's definition: "on-line platforms where ordinary citizens, civic actors, experts, and politicians purposively assemble to provide input, deliberate, inform, and influence policy and decision making" (Tomkova 2009, 2).

The background to the consultation on end-of-life therapies analyzed in this article was the political controversy that arose in 2008 around the expensive anti-cancer drugs Avastin and Erbitux, which were not recommended on cost-effectiveness grounds, but where individual patients had wanted to pay for them privately. The central policy question was whether placing additional value on end-of-life therapies would have been justified, and this was the issue posed in the public consultation by NICE. More than one hundred responses to the consultation came in the form of written submissions via email or regular mail, all of which were available for review electronically on NICE's website. Individual responses, ranging from one to three paragraphs, were then collated by the authors into a single corpus and tagged with the type of respondent: Patient, Carer, Public, Healthcare Professionals (NHSP), Pharmaceutical Companies (PHI), Therapeutics Appraisal Committee Member (TACM), and Other.

Methods of Analysis

Due to the rapid increase in the availability of online documents and Internet content—and thus the emerging need to quickly and effectively interpret these—the automated analysis of large textual corpora with computer assistance has received increasing attention in recent years (see, in particular, Cooley, Mobasher, and Srivastava 1997; Feldman and Sanger 2007; Bauer and Gaskell 2009, Chap. 16, 17). Descending from the older and established tradition of data mining, web mining and text mining have mainly been employed in two separate strands of research. The first has been to discover and process data across thousands and sometimes millions of pages on the World Wide Web. Programs such as WEBSOM (Lagus et al. 1999) and MedMiner, for instance, have been employed to filter and organize large amounts of unstructured information returned by search engines (Tanabe et al. 1999). The second strand of research has been used for more "local" analyses of Internet sources, such as newsgroups, message boards, or electronic brainstorming sessions (Tong and Yager 2006). Although definitions vary, by relying on the use of CATA (Computer Assisted Textual Analysis) and CAQDAS (Computer Assisted Qualitative Data Analysis

Software), text mining can generally be understood as the process of extracting information in large corpora to automatically identify patterns and relationships in textual data (Feldman and Sanger 2007, 17).

A wide variety of text mining software is available on the market today, ranging from those primarily used in qualitative research, integrating such functions as tagging, indexing, and classification (NVivo,[5] MAXQDA,[6] Atlas-ti[7]), to those integrating more quantitative and statistical tools, such as word frequency, cluster analysis, and factorial analysis of the correspondences (WordStat,[8] SimStat,[9] SAS/STAT[10]). The choice to use one type of software or another is thus largely driven by the task at hand, and will vary according to—among other things—the data under consideration, the researcher or analyst's project design and objectives, and the time scale available to carry out the research.

The Alceste[11] software utilized here was originally developed and applied to studies in the humanities (Reinert 1983; 1993). More recently, however, its use has spread to the social sciences (Lahlou 1996; Allum 1998), and it has attracted the attention of political researchers seeking to analyze political speeches (Schonhardt-Bailey 2005), parliamentary debates (Schonhardt-Bailey 2008; Bara, Weale, and Bicquelet 2007), or opinion polls (Brugidou 2003), all of which comprise large, copious amounts of textual data.

Alceste presents several advantages when it comes to analysis of public consultation responses. First, it can handle a large volume of text (indeed, it requires a minimum of 10,000 words to function well). Second, it does not require analysts to create their own dictionaries of key terms, relying instead upon its own vast internal dictionary—it only requires coding variables of interest (in our case, the type of respondent). Third, it has been designed to provide output in the form of cluster and correspondence analyses, particularly useful for the sort of analysis required by consultation responses.

Input into Alceste takes the form of text-file data, stripped of inessential formatting, but with the separate individual responses tagged with information on the type of respondent. The unit of analysis is the sentence (or quasi-sentence),

---

[5] http://www.qsrinternational.com/products_nvivo.aspx.
[6] http://www.maxqda.com/.
[7] http://www.atlasti.com/.
[8] http://www.provalisresearch.com/wordstat/Wordstat.html.
[9] http://www.provalisresearch.com/simstat/simstw.html.
[10] http://www.sas.com/technologies/analytics/statistics/stat/.
[11] ALCESTE stands for *Analyse des Lexèmes Co-occurentsdans les Énnoncés Simples d'un Texte* (Analysis of the co-occurring lexemes within the simple statements of a text). Its algorithm, based on Benzecri's important contributions in textual statistics, was created by Max Reinert at the CNRS. As many other CAQDAS, Alceste is not a free software. It is developed and marketed by the company Image. http://www.image-zafar.com/english/index_alceste.htm.

DOI: 10.2202/1944-2866.1096

also called elementary context units (ECUs). Within these quasi-sentences or ECUs, content words or keywords are automatically identified by the software. The data matrix consists of a set of rows for each sentence, and columns for each word. Within the cells of the matrix, the appearance of the word in the sentence is scored as 1, and 0 otherwise. On this matrix, Alceste computes a descending hierarchical classification of the content words as follows (Guérin-Pace 1998, 79). All sentences are placed together in the same class. That single class is then partitioned into two, according to the criterion of marginal χ^2 values. The initial partitioning aims to maximize the χ^2 values of the margins, dividing the table into two maximally discrete sub-tables. The operation is then repeated until a stable set of partitioned classes is created. The program uses an algorithm to find the table partitioning that maximizes the χ^2 values. Since the classification is purely formal in respect of the numerical entries in the matrix, the analyst is left with the task of interpreting the sense, if any, of the classes that are generated.

The process of descending hierarchical classification establishes the classification of responses, but not the position of respondents within those dimensions. To carry the analysis to this stage, Alceste uses a version of correspondence analysis (Figure 2). In a correspondence analysis, we think of the elements of vocabulary used in a text as defining a multidimensional space, with the distinctive profile of each respondent defining a position in that multidimensional space as well. Different respondents, using different patterns of words, will be defined by these patterns or profiles, and will therefore occupy different positions in the multidimensional space.

In this article, the results generated by the descending hierarchical classification are referred to as the "Standard Analysis." Results output include:

(1) A list of keywords selected for each class.
(2) A cluster analysis showing the degree of association between different classes according to vocabulary and type of respondent.
(3) A correspondence analysis where keywords, classes, and position of respondents are projected on a multidimensional space.
(4) The lists of ECUs automatically selected for each class.

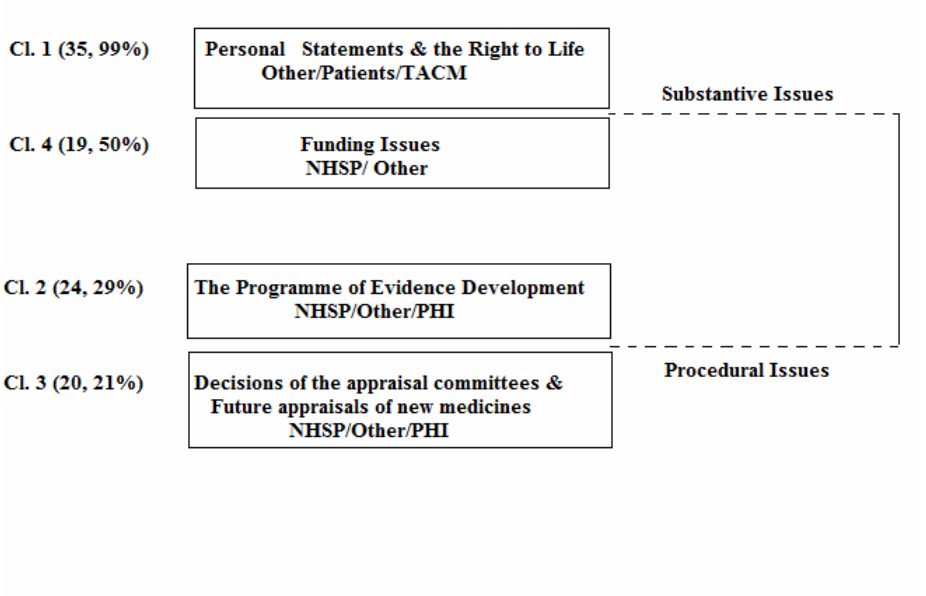
For practical reasons, we only report (2) and (3) here, but (1) will be visible on the correspondence analysis graph (Figure 2) and (4) will be extensively referred to in the Discussion section. We will later introduce the process of "Advanced Analysis" whereby the ECUs selected for a particular class are subjected to a further descending hierarchical classification to refine results.

Results

Standard Analysis

The standard analysis on NICE's consultation on end-of-life medicines yielded four classes divided into two clusters. These are identified in Figure 1, together with the respondents who are statistically associated with each class. For each class it is possible to offer an interpretation by looking at the keywords and sentence segments automatically selected by Alceste or by looking at the correspondence analysis. In what follows we mainly draw on the analysis of the sentence segments (correspondence analysis is presented in Figure 2 for illustrative purposes only).[12]

Figure 1. Standard Analysis Using Alceste: Hierarchical Descending Classification of NICE's Consultation on End-of-Life Medicines (2008).



Note: The figure shows classes 1–4 and the respondent groups statistically associated with them. The percentages correspond to the number of ECUs distributed within each class. With 35.99% of ECUs, class 1 (Personal Statements & the Right to Life) is the biggest class.

[12] The sentence segments referred to and quoted for each class are those with the highest (relative) χ^2. The higher the value of the χ^2, the higher the association between a sentence segment and a class. In other words, sentence segments with the highest χ^2 best capture the vocabulary or theme of a class.

DOI: 10.2202/1944-2866.1096

Two clusters emerged, relating to procedural and substantive issues. The first cluster to emerge from our dataset, comprising classes 1 and 4, pertains to substantive issues raised by the respondents. Class 1 comprises primarily sentence segments from *Patients*, *TA Committee Members*, and *Others*, who relate personal experiences or express subjective views on the availability of (or lack of) life-extending drugs. Within this class of sentences, all respondents tend to stress the "right to life" of patients: for example, "All patients deserve the right to life, and quality of life if they are terminal" or "I have a right to life and if there is a drug to help me in this fight to live, then I deserve to be given it." Moreover, there is a distinctive willingness on the part of the respondents to hasten the decision to make life-extending drugs available; for example: "it seems ridiculous that while this deliberation is taking place, people in urgent need of medicines are becoming more ill by the day" or "The stress and worry you cause by delaying this decision and the pain you place our family members under is totally unacceptable."

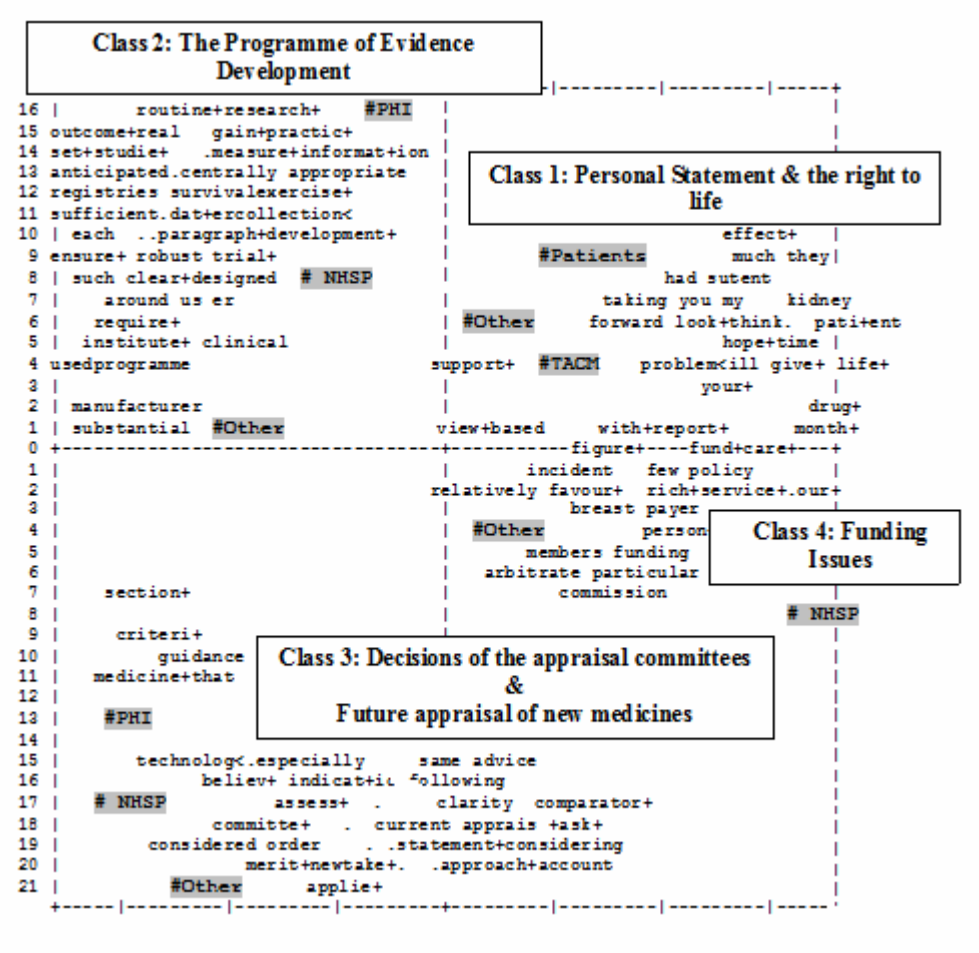
Class 4 comprises mainly the views of *NHS Professionals* and *Others* on funding issues (Figure 1). Although some respondents welcomed a proposal to take end-of-life considerations into special account, a majority of them expressed concerns about the funding of life-extending drugs; for example: "NICE and the NHS will have established a rule of rescue policy that will further distort NHS funding priorities" or "The PCT [Primary Care Trust] is concerned that the proposed changes will result in the funding of medicines to the detriment of other areas of healthcare."

The second cluster to emerge from our analysis comprises classes 2 and 3. In these classes, the issues raised pertain more to the procedural aspect of the consultation rather than to its substance. For instance, in class 2, some respondents pointed out that what NICE considers to be an appropriately designed program of evidence development should be clarified; for example: "we note that acceptance of a medicine under these criteria is to be in the context of an appropriately designed programme of development. While we fully recognize the need to ensure that anticipated survival gains are indeed evident in routine practice we are concerned as to the probability of effectively gathering this data in terms of the timelines for review of the decision."

Similarly, in class 3, some respondents pointed out that decisions made by the appraisal committees concerning the introduction of new medicines and the application of a new higher cost-effectiveness threshold need clarification; for example: "we believe that a number of points in section 2 require clarification. What clarity will NICE be giving to its appraisal committees in the application of new higher cost effectiveness" or "we believe that in final guidance issued for a medicine undergoing this process, it must be made explicitly clear how additional guidance and criteria were applied by the appraisal committee."

These four classes are reported on the factorial analysis of the correspondences (Figure 2) along with the respondents statistically associated with them and the words they most frequently used.

Figure 2. Factorial Analysis of the Correspondences.



Advanced Analysis

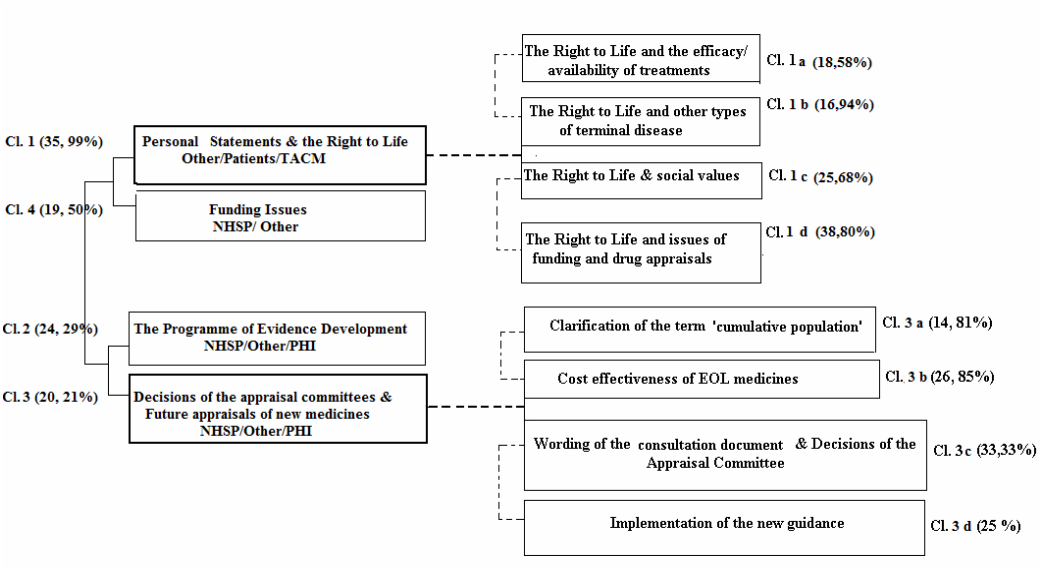
To enhance the analysis of a corpus using Alceste, it is possible to select the ECUs of classes specifically relating to a topic of interest, and then re-analyze those selected sentence segments. The process of classification for this advanced analysis is the same as the standard analysis—hierarchical descending classification—but the focus is on a particular set of words (sentences or quasi-

DOI: 10.2202/1944-2866.1096

sentences) rather than on the whole corpus. The main advantage here is a more refined use of a quantifiable method that retains the qualitative detail of the original source.

For instance, the advanced analysis of class 1 in our dataset (Personal Statements and The Right to Life) yielded four further classes (classes 1a–d). Four types of arguments pertaining to the Right to Life are raised by respondents within these: pragmatic, egalitarian, social, and economic (Figure 3).

Figure 3. Advanced Analyses of Classes 1 and 3.



In class 1a, *Patients* and *Carers* essentially address the issue of the right to life from a perspective of effectiveness at the individual level. They argue that given the seeming effectiveness of life-extending drugs, to deny their availability or access for patients who need them is simply cruel. For example: "I have a very close friend on these medicines. The change in him since taking them is remarkable and is keeping him stable. To deny anybody the opportunity for this treatment is simply wrong. It is a cruel and nasty disease which attacks without prejudice and as such I believe they should continue to be funded for those that need them." Or: "I do not feel qualified to answer the above questions, but I do know that the availability of these types of drugs has enabled a very dear friend of mine to have a longer life than he would have had otherwise."

In class 1b, *Patients*, *Others*, and *Public* respondents emphasize that other types of cancer (in particular, ovarian cancer and kidney cancer) should also be considered terminal diseases. The argument developed in this class is that all

patients must be given every possible care and that they have an equal Right to Life. For instance: "We are concerned particularly with kidney cancer, and we have a close relative with Renal Cell Carcinoma, but the points made below can relate to, and your consultation should consider, other terminal medical conditions and the treatment thereof." Or: "We feel strongly that the NHS must make all drugs that substantially extend life, available in 'end of life' situations, no matter how rare the medical condition and no matter how small the number of actual or potential patients."

Class 1c mainly displays arguments of *Other* and *Public* respondents. In this class, the Right to Life is perceived by respondents as an essential element of a compassionate and fair society. For example, "I agree with the proposition. I suspect society would support the proposition." Or: "I have accompanied terminally ill patients to hospital where they were told 'nothing else can be done for you.' The effect was devastating. Extending life and/or improving the quality of life of terminally ill patients is the action of a caring and compassionate society."

In class 1d, *Patients*, *NHS Professional*, and *Carer* respondents address the Right to Life mainly via economic arguments. For example, "Spend the money please, it isn't yours its mine, and the government has it in its coffers," or: "I have been refused these drugs due to cost effectiveness and this is totally unfair, I have a right to life and if there is a drug to help me in this fight to live then I deserve to be given it. You pay for people to give up smoking, they are a burden on the health costs as they choose to smoke; I did not choose to get cancer."

As with class 1 it was possible to conduct an advanced analysis of class 3 (Decisions of the Appraisal Committees, and Future appraisal of new medicines), which also yielded four classes (3a–d). In these new classes, four types of procedural issues are raised by respondents: the clarification of the term "cumulative population," the cost effectiveness of end-of-life medicines, the wording of the guidance and the decisions of the "Appraisal Committee," and the implementation of the new guidance.

In class 3a, *Pharmaceutical Industries* and *Other* respondents strongly emphasize that the term "cumulative population" needs clarification. For instance: "Despite stating that new indications will be considered on their individual merits, it is unclear from the above statement what the term 'cumulative population' refers to." Or: "Second and subsequent licences for the same product will be considered on their individual merits (will take into account the cumulative population for each product) we would welcome further clarity on how this provision should be interpreted."

In class 3b, *NHS Professionals* and *Other* respondents address the issue of cost effectiveness for end-of-life medicines. For instance: "we recognise that the

DOI: 10.2202/1944-2866.1096

current methodology may not appropriately capture society's values of disutility; an estimated QALY gain of 1 may be viewed differently were this to be obtained through the treatment of an identified patient who would die within 2 years without treatment, or through treating a cohort of 1,000 people of whom an unidentifiable patient is estimated to die, or through treating a cohort of 1000 patients all of whom would benefit from 0.001 QALY."

In class 3c, *Pharmaceutical Industries* and *Other* respondents point out that the wording of the consultation was, at the time, confusing. This, according to them, could affect the decisions of the appraisal committee. For example: "this wording implies that the proposed criteria will be applied retrospectively only after the medicine has undergone appraisal. We are concerned that this may delay the issuing of guidance on new medicines to the NHS and patients unnecessarily, when it may be reasonably obvious at various stages of an appraisal that the medicine will fulfil the eventual criteria."

In class 3d, *NHS Professional* and *Other* respondents discuss the implementation of the new guidance. For instance: "In its own submission to the Richards Review the BMA [British Medical Association] proposed that key elements of the role and function of NICE be re-examined, in particular the speed and transparency of decision-making in respect of its appraisal processes. The current consultation demonstrates a willingness to do so. Given the stated intention to implement this new guidance in January 2009 we would like to see further detail as to what mechanisms are to be put in place for review and feedback in order to assess whether the new methodology is achieving its aims."

Summary of the Results

Overall, the results yielded by the analyses of the public consultation on end-of-life medicines provide support for the utility of the Alceste package for the purposes of textual analysis, such that it might conceivably be of value to policymakers. The software accurately identified the main patterns of argumentation expressed in the consultation, confirmed by manual reading of the textual corpus. It helped to precisely evaluate the association between key themes and respondents, and provided other quantitatively useful statistics, such as frequency and significance (not reported here). Of course, it perhaps comes as no surprise that patients phrase their arguments in terms of the "right to life" whereas pharmaceutical companies are more concerned with the procedural aspects of appraisal when "end of life" is given special weight. Yet, even having an a priori expectation confirmed can be useful to consultation organizers.

Discussion

Like all methodologies, the computer-based application that we have used to analyze NICE's consultation on end-of-life medicines has attracted a certain amount of criticism prompted at times by technical aspects of the software, or by a certain degree of skepticism towards the interpretability of the output (see, for example, Jenny 1999; Laver, Benoit, and Garry 2003). While it is beyond the scope of this paper to address these criticisms, two points for consideration need to be made here. First, different text mining methods are not mutually exclusive; rather they are complementary. It seems very unlikely, and even less desirable, that one of them will ever provide a single solution to the analysis of qualitative or quantitative data. At the very least, exploratory techniques such as the one employed in this article provide a useful first step towards the elaboration of a dictionary-based approach that could be followed by a theory-testing model, or that could be particularly valuable in triangulation exercises to confirm results obtained by other methods. Second, it can legitimately be assumed that categorization, organization, and easy navigation within large corpora can only help in producing more systematic and rigorous analyses of consultation exercises, and will thus contribute to better informed policies by helping, for instance, to identify people's preferences or to establish the profile type of respondents.

In the following discussion, we do not evaluate the performance of the Alceste software in particular; rather we offer an insight into the potential advantages and risks of using text mining methods for the analysis of e-consultations. Starting with the benefits, we outline four advantages: categorization, data reduction, visualization, and speed. We then discuss the inevitable risks and drawbacks of these methods.

Categorization

The problems faced by scholars who work with large text datasets are similar to those of analysts or policymakers who try to make sense of the high volume of disparate responses they receive from consultation exercises. By enabling storage and retrieval of sentence segments, construction of indexes, and cross-referencing, text mining techniques—be they automated (Alceste) or manual (HAMLET and others)—present the advantage of structuring large amounts of qualitative data into predefined or naturally occurring categories (when the algorithm works purely on the basis of co-occurrences).

Retrieval of sentence segments of respondents who share a certain characteristic (e.g., age or profession) or who share similar opinions is in itself a valuable tool. Some software allows further weighting of opinions (see

DOI: 10.2202/1944-2866.1096

MAXQDA 10); others allow one to measure statistically the occurrence of a point of view through frequency and concordance analysis, or comparison of text segments (see, for instance, WordSmith).[13] All these features are particularly useful for the construction of descriptive typologies by type of respondent or point of view expressed.

Data Reduction

Options such as lemmatization (reduction of a word to its root form, also called stemming), the possibility of combining synonyms prior to indexing, and the option offered by many analysis packages to ignore "tool words" such as articles and coordinating conjunctions (see, for instance, WordStat 6.1) provide data-reduction techniques that are particularly useful for the analysis of public consultations. In addition, descending/ascending hierarchical classifications, cluster analysis, and correspondence analysis methods allow the reduction of a large volume of text to its structural components and help to highlight the distinctive points of view that are associated with particular respondents or groups of respondents.

Visualization

Important points are easy to miss when text is read by eye, and the interrelationship between different points of view is hard to identify. Data mining techniques enable the quick generation of visual overviews and mapping of responses, often in several different ways. Dendrograms, 3D scatter plots, heat maps of keyword frequency, and so on make large and complex datasets easier to comprehend. Be they manual or automated, such methods can provide a valuable tool for quick visual identification of the points of view with which respondents are distinctively associated, and for understanding the main dimensions of a public debate.[14]

Speed

Speed of the analysis will be contingent upon whether or not a special dictionary needs to be compiled for the analysis, and on the amount of coding required. Most of the time, however, coding is relatively fast and straightforward. The succinct overview of responses that text mining methods provide could thus enable greater timeliness in the use of consultation responses. UK Cabinet Office guidelines on

[13] http://www.lexically.net/wordsmith/.

[14] See Pieri (2009) for a very judicious use of Atlas-ti to investigate dimensions in the UK debate around the introduction of biometric ID cards in national newspapers.

public consultation currently specify the minimum length of time that consultation involves, but say nothing about how the responses are to be treated. Various improvements can be imagined as a result of greater response speed, including, for example, feedback to participants after a preliminary analysis of initial responses, with the aim of gaining public responses to public responses, thus fostering a more interactive conversation.

However, despite the above advantages, text mining methods present several limitations that have the potential to negatively impact e-consultation analysis. These can be classified into two broad categories: technical and ethical risks.

Technical Risks

Automated text mining techniques, such as the use of the Alceste software, are faster than manual ones, and might therefore appear to be more attractive to consultation organizers or analysts. An important concern about the use of automated text mining methods, however, is that because the classification of responses is automatic, distinctive or marginal points of view may be missed. Precisely because the classification of sentences rests upon the purely formal feature of co-occurrence, it is always possible that sentences may be grouped in a formally correct way, but also in a way such that their meaning is anomalous within the category/class in which they are included. Certain points of view may sometimes be missed entirely by algorithms. There are two principal reasons for this: the sentence segment may overlap several categories and thus belong to a "mixed type" that fits into several lexical categories and addresses different themes, or the sentence segment is too short and uses vocabulary too infrequently to constitute a sentence segment that would be identified by the algorithm.[15]

Another concern is that, while useful tools, lemmatization or stemming (options offered by automated and semi-automated methods) can generate problems in that they can result in important semantic variations being overlooked. For example, in the consultation analyzed here, respondents tended to elaborate on "ills" or "illness." The root form of "ill+" could, a priori, appear to be useful here. Yet, reduction to the root form might actually include different derivative forms such as "ill-defined" or "illdefined," which in this case do not relate to patient illness, but rather to the methods and decisions of health appraisal committees. Reduction to the root form can thus sometimes result in misclassification.

Other technical issues, frequently mentioned in the literature on text mining, that could apply in the case of public e-consultation analysis include the

[15] See Brugidou (2003) on the risks of misclassification by automated techniques.

DOI: 10.2202/1944-2866.1096

danger that analysts distance themselves from the data, especially when attributing numbers to words (Siedel 1991); issues of inter-coder reliability and data preparation (Krippendorff 2004); missing data (Neuendorf 2002); insensitivity to figurative and literal language; and insensitivity to meaning and context, which can result in misclassification when not human-verified (Saldaña 2009; Fielding and Lee 1998).

Ethical Risks

Policy makers or analysts employing text mining methods for e-consultation analysis must consider certain ethical issues in addition to those of informed consent, privacy, and confidentiality. First, respondents may not expect to be research subjects: they may simply be expecting to participate in a general consultation exercise, interacting exclusively with public officials, bodies or organizations running the consultation, not indirectly with an analyst or researcher post hoc; and much less as a specific, traceable data point within a computer-assisted research study. This can be, and has been, a particularly delicate issue for healthcare professionals. Sharf (1999, 247) has described various negative experiences of following up web posts and emailing lists for further research: one woman, on being contacted by a researcher seeking consent to gain insights from breast cancer patients about their personal experiences, became hostile, accusing the researcher of behaving voyeuristically and "taking advantage of people in distress."[16]

A second potential issue concerns statistical interpretation of responses, particularly if analyses are to be returned or made accessible to respondents. Respondents might be confused about or disagree with text mining as a method applied to their answers; indeed, it could be perceived as dehumanizing. In a public consultation, respondents might feel somewhat betrayed that their views and opinions, for which they spent time and effort drafting, eventually resulted in just a dot on a correspondence analysis with no immediate, apparent meaning or import, at least in lay terms. It is, of course, then up to the consultation organizer to clearly and precisely convey the ways in which various qualitative responses can be collated into a quantifiable account of a sample population's views on the matter at hand.

[16] See Esyenbach and Till (2001) for related ethical issues in qualitative research on Internet health communities.

Conclusion

If large-scale e-consultations are still in their formative years, so to speak, it is not unreasonable to believe that they will be increasingly used and that this will require prompt and effective feedback for the organizers. Text mining methods provide valuable assistance for the analysis of unstructured qualitative material and can be gainfully employed to clarify issues and help policy makers reach better decisions. However, such methods also pose important risks that need to be acknowledged. Although technical and ethical issues associated with such methods will not be easily resolved, there are several precautions that could be taken to minimize them. First, consultation organizers should inform respondents about the presence of analysts/researchers among the other professionals in charge of reading responses, obtaining their consent on the types of methods used to analyze the opinions expressed. Second, analysts should verify that the techniques or coding employed will not jeopardize confidentiality or create potential harm to vulnerable groups and/or individuals. Last, but not least, to reduce both technical and ethical risks, researchers need to ensure that the methodology they employ explicitly blends qualitative and quantitative analyses.

While many text mining techniques provide expedient statistical output, the UK Government's prescribed Code of Practice on public consultation is quite explicit on the topic: "The focus should be on the evidence given by consultees to back up their arguments. Analyzing consultation responses is primarily a qualitative rather than a quantitative exercise" (2008, 12).[17] This suggests that the perennial debate between quantitative and qualitative methodologists who try to extract valuable information from texts needs to be updated and better resolved. If the data deluge has just begun, learning how to cope with it will likely be quite a considerable, albeit worthy, challenge for both policymakers and political analysts alike.

References

Allum, N.C. 1998. *A Social Representations Approach to the Comparison of Three Textual Corpora Using Alceste*. London: London School of Economics and Political Science.

Bara, J., A. Weale, and A. Bicquelet. 2007. "Analysing Parliamentary Debate with Computer Assistance." *Swiss Journal of Political Science* 13 (4): 577-605.

[17] http://www.berr.gov.uk/files/file47158.pdf.

DOI: 10.2202/1944-2866.1096

Barber, B.R. 1984. *Strong Democracy: Participatory Politics for a New Age*. Berkeley: University of California Press.

Barnes, M. 1999. "Users as Citizens: Collective Action and the Local Governance of Welfare." *Social Policy and Administration* 33: 73-90.

Bauer, M., and G. Gaskell. 2009. "Computer Assistance." In *Qualitative Researching with Text, Image and Sound*, eds. Bauer, M., and Gaskell, G. London: Sage.

Bicquelet, A. 2009. "On Referendums: A Comparison of French and English Parliamentary Debates Using Computer-assisted Textual Analysis." Government Department. Ph.D. thesis, University of Essex, Wivenhoe.

Brugidou, M. 2003. "Argumentation and Values: An Analysis of Ordinary Political Competence via an Open-Ended Question." *International Journal of Public Opinion Research* 15 (4): 413-430.

Budge, I. 1996. *The New Challenge of Direct Democracy*. Cambridge: Polity Press.

Cooley, R., B. Mobasher, and J. Srivastava. 1997. "Web Mining: Information and Pattern Discovery on the World Wide Web." *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence ICTAI 97*.

Department of Health. 1989. *Working for Patients*, Cmnd55, London: HMSO.

Eysenbach, G., and E.J. Till. 2001. "Ethical Issues in Qualitative Research on Internet Communities." *British Medical Journal* 323: 1103-1105.

Feldman, R., and J. Sanger. 2007. *The Text Mining Handbook: Advanced Approaches to Analyzing Unstructured Data*. Cambridge: Cambridge University Press.

Fielding, N., and Lee, R. 1998. *Computer Analysis and Qualitative Research.* Thousand Oaks, CA: Sage.

Guérin-Pace, F. 1998. "Textual Statistics. An Exploratory Tool for the Social Sciences." *New Methodological Approaches in the Social Sciences* 10 (1): 73-95.

Grossman, K.L. 1995. *The Electronic Republic: Reshaping Democracy in the Information Age*, New York: Viking.

Harrison, S., M. Barnes, and M. Mort. 1997. "Praise and Damnation: Mental Health User Groups and the Construction of Organizational Legitimacy." *Public Policy and Administration* 12 (2): 4-6.

Jenny, J. 1999. "Pour Engager unDébat avec MaxReinert à propos des Fondements Théoriques et des Présupposés des Logiciels d'Analyse Textuelle." *Langage et Société* 90 (1): 73-85.

Kellner, P. 2008. "Expert citizens." *New Statesman*, October 20, 2008, pp. 4-7.

Krippendorff, K. 2004. *Content Analysis: An Introduction to its Methodology*. Thousand Oaks, CA: Sage.

Lagus, K., T. Honkela, S. Kaski, and T. Kohonen. 1999. "Websom for Textual Data Mining." *Artificial Intelligence* 13 (5): 345-364.

Lahlou, L. 1996. "A Method to Extract Social Representations from Linguistic Corpora." *Japanese Journal of Experimental Social Psychology* 35: 278-291.

Laver, M.J., Benoit, K., and Garry, J. 2003. "Extracting Policy Positions fromPolitical Texts Using Words as Data."*American Political Science Review* 97 (2): 311–331.

Margetts, H. 2009. "The Internet and Public Policy." *Policy & Internet* 1 (1). http://www.psocommons.org/policyandinternet/vol1/iss1/art1/

Milewa, T., J. Valentine, and M. Calnan. 1998. "Managerialism and Active Citizenship in Britain's Reformed Health Service." *Social Science and Medicine* 47: 507-517.

McLean, I. 1989. *Democracy and New Technology*. Cambridge: Polity Press.

Needham, C. 2002. "Consultation: A Cure for Local Government?" *Parliamentary Affairs* 55 (4): 699-714.

NHS Community Care Act (NHSCCA). 1990. London: HMSO.

NHS Management Executive (NHSME). 1992. *Local Voices*, Leeds: NHSME.

Neuendorf, K.A. 2002. *The Content Analysis Guidebook*. Thousand Oaks, CA: Sage.

Organisation for Economic Co-operation and Development (OECD). 2001. "Citizens as Partners: Information, Consultation and Public Participation in Policy-making." http://www.ecnl.org/dindocuments/214_OECD_Engaging%20Citizens%20in%20Policy-Making.pdf (accessed September 4, 2010).

Pieri, E. 2009. "ID Cards: A Snapshot of the Debate in the UK Press." *National Centrefor e-Social Science*. Manchester. http://www.ncess.ac.uk/Pieri_idcards_full_report.pdf (accessed September 4, 2010).

Pratchett, L. 1999. "New Fashions in Public Participation: Towards Greater Democracy?" *Parliamentary Affairs* 52 (4): 616-633.

Reddel, T., and G. Woolcock. 2004. "From Consultation to Participatory Governance? A Critical Review of Citizen Engagement Strategies in Queensland." *Australian Journal of Public Administration* 63 (3): 75-87.

Reinert, M. 1983. "Une Méthode de Classification Descendante Hiérarchique: Application a l'Analyse Lexicale par Contexte." *Les Cahiers de l'Analyse desDonnées* 8 (2): 187-198.

Reinert, M. 1993. "Les 'Mondes Lexicaux' et leur 'Logique' a Travers l'Analyse Statistique d'un Corpus de Recits de Cauchemars." *Langage et Société* 66: 5-39.

DOI: 10.2202/1944-2866.1096

Roberts, N. 2004. "Public Deliberation in an Age of Direct Citizen Participation." *The American Review of Public Administration* 34 (4): 315-353.

Rowe, G., and L. Frewer. 2005. "A Typology of Public Engagement Mechanisms." *Science, Technology, and Human Values* 30 (2): 251-290.

Rowe, R., and M. Shepherd. 2002. "Public Participation in the New NHS: No Closer to Citizen Control?" *Social Policy and Administration* 36 (3): 275-290.

Saldaña, J. 2009. *The Coding Manual for Qualitative Researchers*. Thousand Oaks, CA: Sage.

Schonhardt-Bailey, C. 2005. "Measuring Ideas More Effectively: An Analysis of Bush and Kerry's National Security Speeches." *Political Science and Politics* 38 (4): 701-711.

Schonhardt-Bailey, C. 2008. "The Congressional Debate on Partial-Birth Abortion: Constitutional Gravitas and Moral Passion." *British Journal of Political Science* 38 (3): 383-410.

Sharf, B. 1999. "Beyond Netiquette: The Ethics of Doing Naturalistic Discourse Research on the Internet." In *Doing Internet Research*, ed. S. Jones, London: Sage.

Siedel, J. 1991. "Method and Madness in the Application of Computer Technology to Qualitative Data Analysis." In *Using Computers in Qualitative Research.* eds. R.M. Lee, and N.G. Fielding, London: Sage.

Tanabe, L., U. Scherf, L. Smith, J. Lee, L. Hunter, and J. Weinstein. 1999. "'MedMiner: An Internet Text-Mining Tool for Biomedical Information, with Application to Gene Expression Profiling." *Biotechniques* 27 (6): 1210-1217.

Toffler, A., and H. Toffler. 1995. *Creating a New Civilization. The Politics of the Third Wave*. Atlanta: Turner.

Tomkova, J. 2009. "E-consultations: New Tools for Civic Engagement or Facades for Political Correctness?" *European Journal of e-Practice*. http://www.epractice.eu/files/7.4.pdf.

Tong, R., and R. Yager, 2006. "Characterizing Buzz and Sentiment in Internet Sources: Linguistic Summaries and Predictive Behaviors." In *Computing Attitude and Affect in Text: Theory and Applications*, eds. J. Shanahan, Y. Qu, and J. Wiebe, Dordrecht: Springer.

Varian, H. 2010. "Data, Data Everywhere." Special Report on The Data Deluge. *The Economist*, February–March, 1-15.

Walters, L., Aydelotte, J., and J. Miller. 2000. "Putting more Public in Policy Analysis." *Public Administration Review* 60 (4): 349-359.