Computerized Assisted Text Analysis
Course Details

Kenneth Benoit
Methodology Institute
London School of Economics and Political Science
kbenoit@lse.ac.uk

March 23, 2011

Short Outline

The course is intended to survey and characterize methods for systematically extracting information from text for social scientific purposes, starting with classical content analysis methods and proceeding forward to state of the art scaling methods for estimating quantities from text using statistical methods. The course lays a theoretical foundation for text analysis but mainly takes a very practical and applied approach, so that students learn how to apply these methods in actual research. It takes as a starting point more traditional methods of content analysis, but is aimed at the most recent advances in quantitative content analysis that treat words as data to be analysed using statistical tools. The course surveys several of these methods but also applies the statistical framework to more traditional non-automated coding schemes such as the Comparative Manifesto Project and the Policy Agendas Project. It is also designed to cover many fundamental issues such as inter-coder agree- ment, reliability, validation, accuracy, and precision. Lessons will consist of a mixture of theoretical grounding in content analysis approaches and techniques, with hands-on analysis of real texts using content analytic and statistical software.

Prior Knowledge

This course will be designed to provide a first course in applied, computer-assisted text analysis with assuming no prior knowledge of this field.

It will help students, but not be a requirement, to have prior knowledge in the following areas:

- An understanding of basic quantitative methods at the level of an introductory postgraduate social science course.
- Familiarity with the use of some package for working with quantitative data. Ideally this would be Stata or R, but in a pinch, a spreadsheet could be used. Instructional examples will use Stata and R although this only applied to a subset of classes.
- The ability to learn to use text analysis software (on a demonstration basis) such as Wordstat or MaxQDA. Neither of these are required, and no prior knowledge is assumed, but students should be willing and able to try them out. As these are both user-friendly packages and both are available in limited demonstration versions that can be freely downloaded, this should not be too difficult.
- The ability to manipulate text files using a text editor. (It does not matter which text editor you use, but you should use plain text editor — e.g. TextEdit, Notepad, Emacs, BBEdit, etc — and not a word processor such as Microsoft Word.)

Detailed Outline

Meetings

Classes will meet for six sessions of 200 minutes each. Approximately 120 minutes will be lectures, and the other half will consist of "lab" sessions where we will work through exercises in class.

Computer Software

Computer-based exercises will feature prominently in the course, especially in the lab sessions. The use of all software tools will be explained in the sessions, including how to download and install them.

Grading

Grading will be based on a combination of three take-home exercises assigned during the 6-day course, as well as a take-home final exam.

Recommended Texts

The full list of texts is referenced below, but many students will wish to know what texts or software they should consider buying or reading before the course starts. The staple readings (as books) for this course will be Neuendorf (2002) and Krippendorff (2004). Where possible all other readings will be downloadable as pdfs from the course web pages.

Short Course Schedule

	Date	Topic(s)	Details
Thu	14 April	Introduction and Issues in text analysis	Course goals; logistics; software overview; Conceptual foundations; content analysis; objectives; examples.
Mon	18 April	Descriptive inference in text	Co-occurence, concordances, keywords in context; complexity and readability measures; also issues concerning sampling, validity, reliability, agreement in text analysis.
Mon	18 April	Classical quantitative content analysis	Manual unitization and coding approaches, including the CMP, Policy Agendas Project, and self-constructed themes. Software will use MaxQDA.
Tues	19 April	Automated dictionary-based approaches	Dictionary construction, and methods for automatically indexing texts for compiling scales of substantive quantities of interest. Also covers a variety of statistical issues surrounding text types, tokens, and equivalencies, including stemming, lemmatization, and trimming of texts based on word frequencies and *tf-idf*.
Wed	20 April	Words as Data approaches	Automatic "word indexing" and scoring using "Wordscores"; scaling models using parametric (Poisson) and non-parametric (correspondence analysis) methods.
Thur	21 April	Document Scaling	Purely statistical text analysis to recover political ideal points represented in text; some classification methods if we have time; and a brief course review.

Detailed Course Schedule

Day 1: Introduction and Issues in Quantitative Text Analysis

This topic will introduce the goals of the course, the logistics of the course, provide an overview of the topics to be covered, and preview the software to be used. It will also introduce content analysis and quantitative text analysis and discuss how the latter differs from the former. We will cover the conceptual foundations of content analysis and quantitative content analysis, discuss the objectives, the approach to knowledge, and the particular view of texts when performing quantitative analysis. Two examples will be discussed (based on the Gebauer et. al. and Schonhardt-Bailey readings).

Required Reading:

Krippendorff (2004, Ch. 1–3)
Roberts (2000)

Recommended Reading:

(example) Gebauer et al. (2007)
(example) Schonhardt-Bailey (2008)
Neuendorf (2002, Chs. 1–3)

Day 2: Descriptive Inference from Text

This topic covers methods of summmarizing texts and features of texts in order to characterize their properties. It covers many basic quantitative textual measures, including commonly used measures for computing coding reliability. Conceptually it also covers two principal concerns in any systematic text-based analysis: reliability and validity. Other topics to be covered also include sampling concern and choosing and observing units.

Required Reading:

Krippendorff (2004, Chs. 5–6; 11–12)
Banerjee et al. (1999)

Recommended Reading:

Neuendorf (2002, Ch. 4–7)
Benoit et al. (2009)

Assignment:

TBA.

Day 3: Classical Quantitative Content Analysis

Classic (quantitative) content analysis involves the development of coding schemes, the conversion of texts into discrete units and the assignment of codes to each unit based on the coding scheme. This topic covers manual unitization and coding approaches, including the construction of coding frames and different schemes for unitizing texts. It examines two widely used schemes in political

science: the Comparative Manifesto Project and the Policy Agendas Project. User-friendly software packages (e.g. MaxQDA) for applying coding frames will be used for this topic.

Required Reading:

Krippendorff (2004, Chs. 7, 11-12) (CMP) Klingemann et al. (2006, skim but esp. Introduction, Appendixes I–II)

Recommended Reading:

Neuendorf (2002, Chs. 6–7)

Day 4: Automated dictionary-based approaches

Automatic dictionary-based methods involve association of pre-defined word lists with particular quantitative values assigned by the researcher for some characteristic of interest. Here will we will learn methods for constructing dictionaries as well as several methods for using computerized tools to apply the dictionaries to texts. We will also covers a variety of statistical issues surrounding text types, tokens, and equivalencies, including stemming, lemmatization, and trimming of texts based on word frequencies and *tf-idf*.

Required Reading:

Neuendorf (2002, Chs. 6)
Laver and Garry (2000)

Recommended Reading:

Mikhaylov et al. (2010)

Assignment:

TBA.

Day 5: Words as Data approaches

This topic moves beyond human-constructed dictionaries to approaches that dispense althogether with coding frames or dictionaries and instead measure characteristics of the text using relative word frequencies as pure data. In this topic we will introduce the notion of texts as stochastic sources of data, and discuss approaches for making use of this notion. We will cover word frequency distributions, problems and solutions to data sparseness, and related measurement issues that arise using words as data.

Required Reading:

Laver, Benoit and Garry (2003)
Slapin and Proksch (2008)

Recommended Reading:

Benoit, Laver, and Mikhaylov (2009) Monroe and Maeda (2004)

Day 6: Document Scaling

This topic introduces methods for placing documents on continuous dimensions or 'scales'. This topic introduces the major methods for scaling documents and discusses their similarities and differences to other scaling models such as factor analysis and ideal point analysis, and discusses the situations where scaling methods are appropriate.

Required Reading:

Martin and Vanberg (2007)
Benoit and Laver (2007)
Lowe (2008)

Recommended Reading:

Clinton et al. (2004)

Assignment:

TBA

References

Banerjee, M., Capozzoli, M., McSweeney, L., and Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. The Canadian Journal of Statistics/La Revue Canadienne de Statistique, 27(1):3–23.

Benoit, K., Laver, M., and Mikhaylov, S. (2009). "Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions." *American Journal of Political Science* 53(2, April): 495-513

Krippendorff, K. (2004). Content Analysis: An Introduction to Its Methodology. Sage, Thousand Oaks, CA, 2nd edition.

Laver, M., Benoit, K., and Garry, J. (2003). Estimating the policy positions of political actors using words as data. American Political Science Review, 97(2):311–331.

Laver, M. and Garry, J. (2000). Estimating policy positions from political texts. American Journal of Political Science, 44(3):619–634.

Lowe, W. (2008). Understanding Wordscores. Political Analysis, 16(4).

Martin, L. W. and Vanberg, G. (2007). A robust transformation procedure for interpreting political text. Political Analysis, 16(1):93–100.

Mikhaylov, S., Laver, M., and Benoit, K. (2010). Coder reliability and misclassification in comparative manifesto project codings. Paper presented at the 66th MPSA Annual National Conference, Palmer House Hilton Hotel and Towers, April 3–6.

Monroe, B. and Maeda, K. (2004). Talk's cheap: Text-based estimation of rhetorical ideal-points. POLMETH Working Paper.

Neuendorf, K. A. (2002). The Content Analysis Guidebook. Sage, Thousand Oaks, CA.

Roberts, C. W. (2000). A conceptual framework for quantitative text analysis. Quality and Quantity, 34(3):259–274.

Slapin, J. and Proksch, S.-O. (2008). A scaling model for estimating time series policy positions from texts. American Journal of Political Science, 52(8).