ECPR Ljubjana Course 17: Quantitative Text Analysis
Course Details

Kenneth Benoit
Trinity College Dublin
kbenoit@tcd.ie

William Lowe
University of Nottinghamn
will.lowe@nottingham.ac.uk

June 17, 2008

Day to Day Overview

	Topic(s)	**Details**
Day 1 **Mon. 4 August** **(90 minutes)**	Introduction to Quantitative Text Analysis	Goals of course Logistics Topics to be covered Software overview What is QTA and what are its uses, how it differs from non-quantitative content analysis
Day 2 **Tues. 5 August** **(3 hours)**	Issues in Text Analysis	Conceptual foundations Content analysis Objectives Examples
Day 3 **Wed. 6 August** **(3 hours)**	What to analyze?	Sampling concerns Choosing units Texts as stochastic data
Day 4 **Thurs. 7 August** **(3 hours)**	Reliability versus validity	Validity Reliability and agreement Quantitative measures Uncertainty measures
Day 5 **Fri. 8 August** **(3 hours)**	Manual coding: Comparative Manifesto Project as an exercise	Unitizing Coding Strengths and weaknesses of CMP-like approaches

Day 6 **Mon. 11 August** **(3 hours)**	Words as Data	Frequency distributions and sparsity Types, tokens, and linguistic equivalence Beyond English language materials

Day 7 **Tues. 12 August** **(3 hours)**	Classical Content Analysis	Dictionary-based content analysis Constructing a dictionary Measurement issues
Day 8 **Wed. 13 August** **(3 hours)**	Document Classification	Category systems and their applications Overview of classification methods
Day 9 **Thur. 14 August** **(3 hours)**	Document Scaling	Coding, scaling, and categorization Wordscores and Wordfish Text analysis and ideal point analysis
Day 10 **Fri. 15 August** **(3 hours)**	Summary and review of course	

Day 11 **Sat. 16 August**	EXAM

Summary

The course is intended to survey and characterize methods for systematically extracting information from text for social scientific purposes, as well as to teach students how to apply these methods in practical 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 (e.g. Wordscores) but also applies the statistical framework to more traditional non-automated coding schemes (e.g. the Comparative Manifesto Project). It is also designed to cover many fundamental issues such as inter-coder agreement, 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.

Logistics

Meetings. Classes will meet for ten sessions of 3 hours each, approximately half of which will be devoted to exercises in class with the aid of the instructor.

Computer Software. Computer-based exercises will feature prominently in the course, especially in the second half. Software tools will be provided by the instructors and explained in the sessions. In addition, several commercially available software packages will be demonstrated, although their use is not required for this course.

Grading. Grading will be based on a combination of five exercises assigned during the ten-day course, as well as the final exam.

Recommended Texts and Software

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). Many (most) of the other readings will be downloadable as pdfs from the course web page, to be set up by the end of June 2008.

Software will be provided for the students, through a combination of free Stata libraries, free R libraries, and a program called Yoshicoder. We recommend that students attempt to acquire a basic working knowledge of R for the course, and recommend for this purpose the free text *An Introduction to R*.

Detailed Schedule

Introduction to Quantitative Text Analysis — Monday 4 August

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.

Required Reading:
Neuendorf (2002, Chs. 1–2) Roberts (2000)

Recommended Reading:
Krippendorff (2004, Ch. 1)

Assignment: TBA.

Issues in Text Analysis — Tuesday 5 August

In this topic 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. 2–3)
Gebauer, Tang & Baimai (2007)
Schonhardt-Bailey (2008)

Recommended Reading:
Neuendorf (2002, Ch. 3)
Krippendorff (2004, Ch. 4)

Assignment: TBA.

What to Analyze? **Wednesday 6 August**

Topics to be covered include sampling concern and choosing and observing units. It will also introduce the notion of texts as stochastic sources of data, and discuss approaches for making use of this notion.

Required Reading:
Krippendorff (2004, Chs. 5–6)

Recommended Reading:
Neuendorf (2002, Ch. 4)
Benoit, Laver & Mikhaylov (2007)

Assignment: TBA.

Reliability versus Validity **Thursday 7 August**

The two principal concerns in any systematic text-based analysis are reliability and validity, and as suggested by the title, these two goals tend to tradeoff with one another. This topic thoroughly discusses both concepts and discusses their role in designing and evaluating content-analysis based research. This section also covers several key measures of reliability and agreement from a mathematical standpoint.

Required Reading:
Krippendorff (2004, Chs. 11-12)

Recommended Reading:
Neuendorf (2002, Chs. 6–7)
Banerjee, Capozzoli, McSweeney & Sinha (1999)
Mikhaylov, Laver & Benoit (2008)

Assignment: TBA.

Manual Coding Approaches **Friday 8 August**

Manual coding schemes involve the conversion of texts into discrete units and the assignment of codes to each unit based on a pre-defined scheme. Here we discuss discuss this generally and then apply it to the longest-running scheme in political analysis, the Comparative Manifesto Project.

Required Reading:
Krippendorff (2004, Ch. 7)
Klingemann, Volkens, Bara, Budge & McDonald (2006, skim but esp. Introduction, Appendixes I–II)

Recommended Reading:
Neuendorf (2002, Chs. 6)
Mikhaylov, Laver & Benoit (2008)

Assignment: TBA.

Words as Data — Monday 11 August

Words and their frequencies in text have rather different statistical properties to many other types of variable used in quantitative analyses. This topic provides an overview and practical investigation into word frequency distributions, problems and solutions to data sparseness, and related measurement issues that arise using words as data. This data-oriented topic provides the grounding for models of the relationship between individual word occurrences and more substantively interesting quantities. It also discusses how to make best use of non-english language materials.

Required Reading:
Krippendorff (2004, Ch. 12)

Recommended Reading:
Neuendorf (2002, Resource 3)

Assignment: TBA.

Classical Content Analysis — Tuesday 12 August

Traditionally, content analyses have rested on the application of a manually-constructed content dictionary to texts and the analysis of the word frequency counts it generates. The topic introduces this methodology in the context of an overarching theoretical framework of quantitative text analysis models. Students will be introduced to some of the currently available software and will use it to replicate published analyses.

Required Reading:
Neuendorf (2002, Chs. 6)
Alexa & Zuell (2000)
Yoshikoder software: http://www.yoshikoder.org

Recommended Reading:
Laver & Garry (2000)

Assignment: TBA.

Document Scaling — Wednesday 13 August

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:
Laver, Benoit & Garry (2003)
Slapin & Proksch (2008)
Wordscores http://wordscores.com, Wordfish: http://wordfish.org.

Recommended Reading:
Benoit & Laver (2003)
Proksch & Slapin (2008)

Assignment: TBA.

Document Classification **Thursday 14 August**

This topic discusses statistical methods for classifying documents into categories, the nature of category systems, and special issues arising from using words as data. The topic also introduces validation and reporting methods for classifiers and discusses where these methods are applicable.

Required Reading:
TBA.

Assignment: TBA.

Summary and review **Friday. 15 August**

References

Alexa, Melina & Cornelia Zuell. 2000. "Text Analysis Software: Commonalities, Differences and Limitations: The Results of a Review." *Quality and Quantity* 34(3):299–321.

Banerjee, M., M. Capozzoli, L. McSweeney & D. Sinha. 1999. "Beyond Kappa: A Review of Interrater Agreement Measures." *The Canadian Journal of Statistics/La Revue Canadienne de Statistique* 27(1):3–23.

Benoit, Kenneth & Michael Laver. 2003. "Extracting Policy Positions From Political Texts Using Phrases As Data: A Research Note." Paper presented the 2003 annual meeting of the Midwest Political Science Association, Palmer House Hilton and Towers, Chicago, IL, 3–6 April.

Benoit, Kenneth, Michael Laver & Slava Mikhaylov. 2007. "Estimating Party Policy Positions with Uncertainty Based on Manifesto Codings." Presented at the 2007 Annual Meeting of the American Political Science Association, Hyatt Regency and Sheraton Chicago, Chicago, Illinois, August 30–September 2, 2007.

Gebauer, Judith, Ya Tang & Chaiwat Baimai. 2007. "User requirements of mobile technology: results from a content analysis of user reviews." *Information Systems and E-Business Management* (7 December 2007).

Klingemann, Hans-Dieter, Andrea Volkens, Judith Bara, Ian Budge & Michael McDonald. 2006. *Mapping Policy Preferences II: Estimates for Parties, Electors, and Governments in Eastern Europe, European Union and OECD 1990-2003*. Oxford: Oxford University Press.

Krippendorff, Klaus. 2004. *Content Analysis: An Introduction to Its Methodology*. 2nd ed. Thousand Oaks, CA: Sage.

Laver, Michael & John Garry. 2000. "Estimating Policy Positions from Political Texts." *American Journal of Political Science* 44(3):619–634.

Laver, Michael, Kenneth Benoit & John Garry. 2003. "Estimating the policy positions of political actors using words as data." *American Political Science Review* 97(2):311–331.

Mikhaylov, Slava, Michael Laver & Kenneth Benoit. 2008. "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.

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

Proksch, Sven-Oliver & Jonathan Slapin. 2008. "Position-Taking in European Parliament Speeches." Paper presented at the Annual Meeting of the Midwest Political Science Association, March 2008.

Roberts, Carl W. 2000. "A Conceptual Framework for Quantitative Text Analysis." *Quality and Quantity* 34(3, August):259–274.

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

Slapin, Jonathan & Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time Series Policy Positions from Texts." *American Journal of Political Science* 52(8).