

# ECPR Ljubjana Course 17: Quantitative Text Analysis

## Course Details

Kenneth Benoit  
Trinity College Dublin  
[kbenoit@tcd.ie](mailto:kbenoit@tcd.ie)

Will Lowe  
University of Maastricht  
[w.lowe@maastrichtuniversity.nl](mailto:w.lowe@maastrichtuniversity.nl)

August 2, 2009

### Short Outline

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 but also applies the statistical framework to more traditional non-automated coding schemes such as 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.

### Prior Knowledge

Ideally, students in this course will have prior knowledge in the following areas:

- A basic understanding of probability and statistics at the level of an introductory postgraduate social science course. Understanding of regression analysis is presumed;
- Familiarity with the language R.  
We recommend that students attempt to acquire a basic working knowledge of R for the course, and recommend for this purpose the free electronic text: *An Introduction to R*;
- The ability to manipulate text files using a text editor.

### Detailed Outline

#### Meetings

Classes will meet for ten sessions of 3 hours each, approximately one third of which will be devoted to exercises in class with the aid of the instructors and teaching assistant.

#### 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 also 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

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
Mon.	3 Aug.	Introduction	Course goals; logistics; software overview
Tues.	4 Aug.	Issues in text analysis	Conceptual foundations; content analysis; objectives; examples
Wed.	5 Aug.	What to analyze?	Sampling concerns; choosing units; texts as stochastic data
Thur.	6 Aug.	Reliability versus validity	Validity; reliability and agreement; quantitative measures; uncertainty measures
Fri.	7 Aug.	Manual coding: The Comparative Manifesto Project	Unitizing; coding; strengths and weaknesses of CMP-like approaches
Mon.	10 Aug.	Words as Data	Frequency distributions and sparsity; Types, tokens, and equivalences; non-English language material
Tues.	11 Aug.	Classical Content Analysis	Dictionary-based content analysis; constructing a dictionary; measurement issues
Wed.	12 Aug.	Document Classification	Category systems and applications; overview of classification methods
Thur.	13 Aug.	Document Scaling	Coding, categorization, and scaling; Wordscores and Wordfish; Text analysis and ideal points
Fri.	14 Aug.	Summary and Course Review	
Sat.	15 Aug.	EXAM	

## **Detailed Course Schedule**

### **Day 1: Introduction to 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.

#### **Required Reading:**

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

#### **Recommended Reading:**

Krippendorff (2004, Ch. 1)

#### **Assignment:**

TBA.

### **Day 2: Issues in Text Analysis**

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 et al. (2007)  
Schonhardt-Bailey (2008)

#### **Recommended Reading:**

Neuendorf (2002, Ch. 3)  
Krippendorff (2004, Ch. 4)

#### **Assignment:**

TBA.

### **Day 3: What to Analyze?**

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 et al. (2007)

**Assignment:**

TBA.

**Day 4: Reliability versus Validity**

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 et al. (1999)  
Mikhaylov et al. (2008)

**Assignment:**

TBA.

**Day 5: Manual Coding Approaches**

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 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 et al. (2006, skim but esp. Introduction, Appendixes I–II)

**Recommended Reading:**

Neuendorf (2002, Chs. 6)  
Mikhaylov et al. (2008)

**Assignment:**

TBA.

## **Day 6: Words as Data**

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.

### **Required Reading:**

Krippendorff (2004, Ch. 12)

### **Recommended Reading:**

Neuendorf (2002, Resource 3)

### **Assignment:**

TBA.

## **Day 7: Classical Content Analysis**

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)

Laver and Garry (2000)

Alexa and Zuell (2000a)

### **Recommended Reading:**

Alexa and Zuell (2000b)

Bara et al. (2007)

Pennebaker and Chung (2008)

Schonhardt-Bailey (2005)

### **Assignment:**

TBA.

## **Day 8: Document Classification**

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:**

Hilliard et al. (2006)  
McIntosh et al. (2007)

**Recommended Reading:**

Hopkins and King (2007)  
Quinn et al. (2006)  
Yu et al. (2008)

**Assignment:**

TBA.

**Day 9: 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:**

Laver et al. (2003)  
Slapin and Proksch (2008)  
Lowe (2008)

**Recommended Reading:**

Clinton et al. (2004)  
Benoit and Laver (2003)  
Proksch and Slapin (2008)  
Martin and Vanberg (2007)  
Monroe and Maeda (2004)

**Assignment:**

TBA

## References

- Alexa, M. and Zuell, C. (2000a). Text analysis software: Commonalities, differences and limitations: The results of a review. *Quality and Quantity*, 34(3):299–321.
- Alexa, M. and Zuell, C. (2000b). Text analysis software: commonalities, differences and limitations: the results of a review. *Quantity and Quality*, 34:299–321.
- 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.
- Bara, J., Weale, A., and Biquelet, A. (2007). Analysing parliamentary debate with computer assistance. *Swiss Political Science Review*, 13(4):577–605.
- Benoit, K. and Laver, M. (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, K., Laver, M., and Mikhaylov, S. (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.
- Clinton, J., Jackman, S., and Rivers, D. (2004). The statistical analysis of roll call voting: A unified approach. *American Journal of Political Science*, 98(2):355–370.
- Gebauer, J., Tang, Y., and Baimai, C. (2007). User requirements of mobile technology: Results from a content analysis of user reviews. *Information Systems and E-Business Management*.
- Hilliard, D., Purpura, S. J., and Wilkerson, S. (2006). Computer assisted topic classification for mixed methods social science research. *Journal of Information Technology and Politics*, 4(4).
- Hopkins, D. and King, G. (2007). Extracting systematic social science meaning from text. Paper presented at MPSA 2007, Chicago IL.
- Klingemann, H.-D., Volkens, A., Bara, J., Budge, I., and McDonald, M. (2006). *Mapping Policy Preferences II: Estimates for Parties, Electors, and Governments in Eastern Europe, European Union and OECD 1990-2003*. Oxford University Press, Oxford.
- 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.
- McIntosh, W., Evans, M., Lin, J., and Cates, C. (2007). Recounting the courts? applying automated content analysis to enhance empirical legal research. *Journal of Empirical Legal Studies*, 4(4):1041–1057.
- Mikhaylov, S., Laver, M., and Benoit, K. (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.

- 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.
- Pennebaker, J. W. and Chung, C. K. (2008). Computerized text analysis of al-Qaeda transcripts. In Krippendorf, K. and Bock, M. A., editors, *The Content Analysis Reader*. Sage.
- Proksch, S.-O. and Slapin, J. (2008). Position-taking in european parliament speeches. Paper presented at the Annual Meeting of the Midwest Political Science Association, March 2008.
- Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M., and Radev, D. R. (2006). An automated method of topic-coding legislative speech over time with application to the 105th-108th U.S. senate. Paper presented to the Society for Political Methodology. University of California at Davis.
- Roberts, C. W. (2000). A conceptual framework for quantitative text analysis. *Quality and Quantity*, 34(3):259–274.
- Schonhardt-Bailey, C. (2005). Measuring ideas more effectively: An analysis of Bush and Kerry's national security speeches. *PS: Political Science and Politics*, 38.
- Schonhardt-Bailey, C. (2008). The congressional debate on partial-birth abortion: Constitutional gravitas and moral passion. *British Journal of Political Science*, 38:383–410.
- 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).
- Yu, B., Kaufmann, S., and Diermeier, D. (2008). Classifying party affiliation from political speech. *Journal of Information Technology and Politics*, 5(1):33–48.