10th ECPR Summer School in Methods and Techniques, 23 July - 8 August
University of Ljubljana, Slovenia
Course Description Form[1] - Week 2 course, 15 hrs (3 – 7 August 2015)

Course title

SD103. Content Analysis

Instructor details

First name, last name: Kostas Gemenis
Department/Unit: Department of Public Administration
Institution: University of Twente
Full postal address for ECPR correspondence: Department of Public Administration, University of Twente, P.O. Box 217, 7500 AE, Enschede, The Netherlands
Phone: 0031 53 489 3256
Fax: 0031 53 489 2159
E-mail : k.gemenis@utwente.nl

Short Bio

Kostas Gemenis is Assistant Professor of Research Methods at the Department of Public Administration, University of Twente. His research interest include measurement in the social sciences, and content analysis with applications to estimating the policy positions of political actors. He is also currently involved in Preference Matcher <www.preferencematcher.org>, a consortium involving researchers who collaborate in developing e-literacy tools designed to enhance voter education. http://www.utwente.nl/mb/pa/staff/gemenis/

Prerequisite knowledge

Participants are expected to be familiar with basic statistical concepts such as measures of central tendency (mean, median), dispersion (standard deviation), tests of association (Pearson's r) and inference (χ^2, t-test). These material are covered in the first few chapters of introductory statistics or data analysis textbooks. A useful example is Pollock P.H. III, *The Essentials of Political Analysis*, fourth edition (Washington, DC: CQ Press, 2012), Chapters 2, 3, 6, and 7. Some familiarity with SPSS, and Stata or R statistical software is also desirable.

Short course outline

The course will introduce participants to the family of methods known as 'content analysis' using a variety of examples from political and other disciplines. The course will cover the basic aspects of content analysis relating to creating coding schemes, document selection, coding, and scaling. Particular attention will be paid to evaluating measurement in the context of content analysis in terms of reliability and validity. The course will cover different approaches to both manual and computer-assisted coding in content analysis, and will be taught in a mix of lectures and seminars. After the lectures, participants will be asked to do some hands-on content analysis exercises (codebooks, text material, and software where needed will be provided via Moodle) and the results will be discussed during the seminars. In additions, participants will be able to present their own project in class and get feedback from the instructors and the other participants.

---

[1] *Disclaimer: the information contained in this course description form may be subject to subsequent adaptations (e.g. taking into account new developments in the field, specific participant demands, group size etc.). Registered participants will be informed in due time in case of adaptations.*

Long course outline

Content analysis is typically defined as a method whose goal is to summarize a body of information, often in the form of text, and to make inferences about the actor behind this body of information. This implies that content analysis can be seen as a data reduction method since its goal is to reduce the text material in to more manageable bits of information. As these manageable bits are often in the form of quantitative data (i.e. numbers), most often than not, researchers refer to content analysis as a 'quantitative' method. Content analysis can be also seen as a method for descriptive inference. Weber (1990, p. 9) for instance, defines content analysis as 'a method that uses a set of procedures to make valid inferences from text'. The idea is that, by analysing the textual output of an actor, we can infer something about this actor. This conceptualization of content analysis implies that we can use it as a tool for measurement in the social sciences.

Since many social science concepts are not directly observable, content analysis can provide a useful method in which we can measure quantities of interest that are otherwise difficult to estimate. For instance, by content analysing the speeches of a political leader, we can infer whether this leader is charismatic, populist, authoritarian, liberal, and so on. By content analysing the content of newspaper editorials, we can infer whether there has been media bias for a particular candidate during an election campaign. This view of content analysis, however, assumes that we are employing the scientific method and therefore any content analysis application should be concerned with replicability, objectivity, reliability, validity and so on (Neuendorf 2002, pp. 10-15). As such, content analysis should *not* be confused with other approaches/methods in the 'qualitative' research tradition such as discourse analysis, rhetorical analysis, constructivism, ethnography and so on. Having said that, it should not assume that content analysis is a strictly 'quantitative' method. Content analysis involves the use of statistics, but much of what goes on can be considered to be of 'qualitative' nature. Our reading of texts is qualitative and so is the assigning of sentences or words in the categories of a coding scheme. Similarly, the construction of a dictionary for computer-assisted content analysis involves decisions that have a great deal of subjectivity. The purpose of content analysis is to reduce this subjectivity and make the results as much replicable/objective as possible. To do so, we will employ some statistical concepts, but will not engage in sophisticated statistical analyses. Content analysis is a fairly simple method to employ, but should not be done in an overly simplistic way. In this respect, the course will use many examples to illustrate the promises of content analysis in the social sciences, but also the pitfalls in terms of design misspecification or sloppy implementation.

Using the chapters in Krippenforff (2004) and Neuendorf (2002), the course will introduce participants to the basic concepts and building blocks in content analysis designs. The course will also look at the, often overlooked, distinction between the analysis of manifest content and judgemental coding. The course will also contrasting manual to computer-assisted content analysis. In addition, the course will compare manual to computer-assisted content analysis. Finally, the course will look at relationship between reliability and validity and outline methods for estimating inter-coder reliability and validating the results produced by computer-assisted content analysis. The following issues, will be addressed (among others):

- Coding scheme (What are the theoretical underpinnings of the coding scheme? How are the categories selected and operationalized? What are the coding units? How is coding performed? Is our coding scheme valid?)

- Selection of documents (What guides the selection of texts? Are texts sufficiently comparable? Are our documents valid and reliable indicators of the quantities of interest?)
- Aggregation (Are texts coded by different coders? If so, how are their results aggregated? If not, how can we ensure inter-coder reliability? Is our coding reliable?)
- Scaling (Are we estimating the quantities of interest directly? If not, how do we scale data in order to estimate the quantities of interest? Is our scaling valid and reliable?)

The format of the course will be a mixture of lectures, seminars, participant assignments and presentations. The lectures will outline the building blocks, challenges and trade-offs in content analysis. Participants will need to complete two assignments on manual coding that will be discussed and extended during seminars. The seminar on computer-assisted content analysis will be a hands-on session with different methods and software. Finally, participants will be able to present their own content analysis project and receive feedback from the instructor and other participants.

The lectures, seminars, and assignments will use examples from different areas and disciplines in the social sciences such as sentiment analysis of the press, frames analysis of social movements, estimating the positions of political actors, and agenda-setting in the EU.

Day-to-day schedule (Monday 3 August - Friday 7 August)

	Topic(s)	**Details**
Day 1	Introduction to content analysis; Manual content analysis I	Lecture (90 mins.) • Brief presentation of participants and their research projects; Defining content analysis; Key concepts in content analysis; Reliability and validity and their relationship to measurement error; Manifest versus judgemental coding; Manual versus computer-assisted coding Lecture (90 mins.) • Designing a manual content analysis coding scheme; Selecting the appropriate documents; Coding the documents; Inter-coder reliability; Scaling the coded data. • Assignment: coding a European Commission work programme using a manual coding scheme
Day 2	Manual content analysis II	Seminar (90 mins.) • Discussion of the coding assignment; Calculating inter-coder reliability using Krippendorff's alpha; Fine-tuning manual coding schemes Lecture (90 mins.)

		• Judgemental coding; Designing a judgemental coding scheme; Perceptual agreement; Iterative methods; Anchoring vignettes • Assignment: coding protest pamphlets using a judgemental coding scheme
Day 3	Computer-assisted content analysis I	Seminar (90 mins.) • Discussion of the coding assignment; Calculating perceptual agreement using Van der Eijk's A; Fine-tuning judgemental coding schemes Lecture (90 mins.) • The promises of computer-assisted content analysis; Selecting, cleaning and formatting documents; Constructing a coding dictionary
Day 4	Computer-assisted content analysis II	Lecture (90 mins.) • Scaling models in computer-assisted content analysis and their assumptions; Supervised and unsupervised methods; Wordscores and Wordfish; Validating computer-assisted content analysis Seminar (90 mins.) • Sentiment analysis in print media using a computer coding dictionary; Scaling parties' positions in party manifestos using Wordscores
Day 5		Presentations (90 mins.) • Presentations of participants projects and discussion Lecture (90 mins.) • Comparisons among the different approaches to content analysis; trade-offs in reliability and validity; crowd-sourcing manual coding

Day-to-day reading list

	Readings
Day 1	Compulsory Neuendorf, Kimberly A. (2002) *The content analysis guidebook*. Thousand Oaks, CA: Sage, Chapter 1 (defining content analysis) Krippendorff, Klaus (2004) *Content analysis: An introduction to its methodology*, second edition. Thousand Oaks, CA: Sage, Chapters 5 (unitizing) and 7 (coding) Riffe, Daniel, Stephen Lacy, and Frederick G. Fico (2005) *Analyzing media messages: Using quantitative content analysis in research*. Mahwah, NJ: Lawrence Erlbaum, Chapter 6 (reliability) Optional Budge, Ian (2001) Validating party policy placements. *British Journal of Political Science* 31: 210–223. Gemenis, Kostas (2013) What to do (and not to do) with the Comparative Manifestos Project data. *Political Studies* 61(S1): 3–23
Day 2	Hayes, Andrew F., and Klaus Krippendorff (2007) Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures* 1: 77–89. Dalton, Russell J., Paul A. Beck, and Robert Huckfeldt (1998) Partisan cues and the media: Information flows in the 1992 presidential election. *American Political Science Review* 96: 111–126. Gemenis, Kostas (2014) An iterative expert survey approach for estimating parties' policy positions. *Quality & Quantity* , doi: 10.1007/s11135-014-0109-5
Day 3	Grimmer, Justin, and Brandon M. Stewart (2013) Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21: 267–297. Laver, Michael, and John Garry (2000) Estimating policy positions from political texts. *American Journal of Political Science* 44: 619–634. Adcock, Robert, and David Collier (2001) Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review* 95: 529-546.

Day 4	Hopkins, Daniel J., and Gary King (2010) A method of automated nonparametric content analysis for social science. *American Journal of Political Science* 54: 229-247. Laver, Michael, Kenneth Benoit, and John Garry (2003) Extracting policy positions from political texts using words as data. *American Political Science Review* 97: 311–331.
Day 5	Benoit, Kenneth, Drew Conway, Benjamin E. Lauderdale, Michale Laver, and Slava Mikhaylov (2014) Crowd-sourced coding of political texts. Unpublished manuscript.

Software and hardware requirements

Software programme
Stata version 11 or higher, with wordscores, concord, agrm, polychoric user-written packages installed
SPSS version 20 or higher, with KALPHA user-written menu installed
Yoshikoder free software download
Lexicoder free software download

Hardware requirements
N/a

Literature
N/a

Lecture room requirement
Classical lecture style room for lectures, computer room for seminars.

Preferred time slots
Morning.

Other recommended courses (before or after this course)

Before this course:

	Course title	Summer School	Winter School
1	Research Designs	X	
2	Qualitative Data Analysis: Methods and Procedures	X	
3	Basics of Inferential Statistics for Political Scientists	X	
4	Research design fundamentals		X
5			

After this course:

	Course title	Summer School	Winter School
1			
2			
3			
4			
5			

as clear as you can to give indications about what participants can expect.