Things to Do with Words
Content Analysis

Contents

Content analysis: A quantitative technique

Content analysis is a text analysis technique for the social sciences. Developed by Harold D. Lasswell in the early 1940s, its goal was to gain objective and systematic insights into the content of a set of documents. And the mean to that goal was quantification: counting specific themes/symbols of interest to the analyst. Those original goal and mean are still with us today.

Lasswell looked at writing for what it would tell him about propaganda and communication – propaganda as “the control of opinion by significant symbols… [and by] forms of social communication”; and communication as answers to the questions “Who, says what, how, to whom, with what effect?”, the “what” question being the primary concern of content analysis (Lasswell 1926:11, see also 1927:627; Lasswell et al. 1952:12).

And building upon Lasswell’s earlier attempts to quantify through counts, the technique was born as quantitative. As Lasswell would later put it: “There is clearly no reason for content analysis unless the question one wants answered is quantitative.” (Lasswell et al. 1952:45) Lasswell proposed to quantify by counting words, themes, symbols (“a technical term for words”, Lasswell et al. 1952:29), symbol clusters, but also concepts and ideas (Lasswell et al. 1952:29, 27, 34, 54, 69).

While avoiding specifics, Lasswell does make two general recommendations for the design of a coding scheme:

1. stay away from coding categories that are too abstract (a recommendation all too often ignored in content analysis projects). Reliability increases with “coding by explicit symbol rather than by more interpretive categories.” (Lasswell et al. 1952:62)
2. resist temptation of “attempting to reproduce all the possible complexities of language” (Lasswell et al. 1952:52). Simplify, “painful” as simplification may be. Or… You may end up with too much (costly) data only a fraction of which you will ever use or data that will require “a content analysis of one’s collected content-analysis data.” (Lasswell et al. 1952:52)

With that in mind, what do you want to know about Mrs. Felton’s letter of *The Atlanta Journal*? Southern attitudes about race (and gender) relations in Jim Crow South? About lynching? The role (failure) of religion in instilling morality into individuals? The role (failure) of the courts in upholding the law? The design of coding categories does indeed reflect the questions asked of texts.

But what is content analysis? Early definitions are clear enough (e.g., Shapiro and Markoff, 1997):

The technique known as content analysis … attempts to characterize the meanings in a given body of discourse in a systematic and quantitative fashion. (Kaplan, 1943: 230)

Content analysis is a research technique for the objective, systematic, and quantitative description of the manifest content of communication. (Berelson, 1952: 18)

'Content analysis' … refer[s] to the objective, systematic, and quantitative description of any symbolic behavior. (Cartwright, 1953: 424)

And it was Lasswell who in 1949 wrote a famous article titled 'Why be quantitative?'. In that paper, Lasswell forcefully argued the case for a quantitative approach in the study of politics.

The **call to arms**

Lasswell (1949: 52) concluded: 'Why, then, be quantitative about communication? Because of the scientific and policy gains that can come of it.'

In one of the early textbooks, Holsti (1969: 14) specifically wrote: 'Content analysis is any technique for making inferences by objectively and systematically identifying specified characteristics of messages. … Our definition does not include any reference to quantification'. Stone, the developer of The General Inquirer, one of the early attempts at computer understanding of natural languages, similarly wrote: 'Content analysis is any research technique for making inferences by systematically and objectively identifying specified characteristics within text' (in Stone et al., 1966: 5). And Krippendorf's classic textbook, Content Analysis: An Introduction to Its Methodology,

One of the most common approaches to content analysis is *thematic analysis*, where the coding scheme is based on categories designed to capture the dominant themes present in a text.

Thematic Content Analysis. Thematic analysis is the most common approach in content analysis. In thematic analysis the coding scheme is based on categories designed to capture the dominant themes in a text. Unfortunately there cannot be a universal coding scheme: different texts emphasize different things; and different investigators could be looking for different things in the same texts. That is why the development of good thematic analysis requires intimate familiarity with the input text and its characteristics. It also requires extensive pre-testing of the coding scheme.

Referential Content Analysis. Thematic content analysis is a good tool for teasing out the main themes expressed in a text. But meaning is also the result of other kinds of language games: backgrounding and foregrounding information, silence and emphasis, different ways of describing the same thing. Referential content analysis is a tool better suited than thematic analysis to capture the complexity of language in the production of meaning. Referential analysis, Krippendorf tells us, is used when “the task is to ascertain how an existing phenomenon is portrayed” (Krippendorf, 1980, p. 62). And that portrayal is affected by the choice of nouns and adjectives, or even of different syntactical constructs, such as passive or active forms (Franzosi, in press).

“Story Grammars” and the Structure of Narrative. From the very early days of the technique researchers have pleaded for an approach to content analysis grounded on the linguistic properties of the text (e.g. de Sola Pool, Hays, Markoff, Shapiro, and Weitman; see Franzosi, chapter 1, in press). In recent years, several authors have attempted to bridge the gap between linguistics and content analysis proposing some basic variants of “story grammars” (or text grammars or semantic grammars, as they are also known) for the analysis of narrative texts (Abell, 1987; Shapiro and Markoff, 1998; Franzosi, in press). Basically, a story grammar is nothing but the simple structure of the 5 Ws of journalism: who, what, where, when, and why (and how); someone doing something, pro or against someone else, in time and space; or subject-action-object or, better, agent-action-patient/beneficiary and their modifiers (e.g. type and number of subjects and objects; time and space, reasons and outcomes of actions). The technique has been used to great effect in the study of social and protest movements.

How Content Analysis Works

Quantification may not necessarily be part of any formal definition of content analysis but, no doubt, what the different approaches to content analysis have in common is a concern with numbers – which distinguishes them from typically qualitative approaches to the analysis of texts and symbolic material, such as NARRATIVE ANALYSIS, CONVERSATIONAL ANALYSIS, DISCOURSE ANALYSIS or SEMIOTICS. But, if so, how does content analysis transform words into numbers? The engine of this transformation is certainly the “coding scheme,” although other “scientific” procedures play a vital role in the process: SAMPLING, VALIDITY and RELIABILITY.

Coding scheme. The coding scheme is the set of all coding categories applied to a collection of texts – where a “coding category” identifies each characteristic of interest to an investigator. The scheme is systematically applied to all selected texts for the purpose of extracting uniform and standardized data: if a text contains information on any of the coding categories of the coding scheme, the relevant coding category are “ticked off” by a human coder (a process known as CODING in content analysis; modern qualitative software such as N6 – ex NUD*IST – or ATLAS.ti allow users to code text directly on the computer and to make up coding categories as they go along). When coding is completed, the ticks are added up for each coding category. In the end, the alchemic recipe of content analysis for getting numbers out of words is simple: you count. Whether it is words, themes, references, actors and their actions, depending upon the specific type of content analysis, the numbers are the result of counting.

Sampling. Sampling provides an efficient and cost-effective way to achieve research results. Rather than working with all the possible available sources and documents, investigators may decide to sample an appropriate number of sources and an appropriate number of documents. Sampling may be particularly appealing in content analysis given that it is a very labor-intensive, and therefore expensive, technique (and sampling can also be used to check the reliability of coded data through acceptance sampling schemes, Franzosi, in press). By sampling, you may focus on a subset of newspapers, letters, or other documents of interest. You may focus on selected years. And even for those years, you may read one or more days a week of a newspaper, one or more weeks a month, one or more months a year. Setting up a sampling frame that will not bias the data requires a sound knowledge of the input material. For example, relying on only the Sunday issues of a newspaper, how would that affect data? To answer this and similar questions, any choice of sources and of sampling frame should be based on systematic comparative analyses. The validity of your data (i.e., the basic correspondence between a concept that you are trying to measure and your actual measurements) will be a function of the sources and sampling frame adopted.

Inter-coder reliability. In content analysis issues of reliability (i.e., of the repeatability of measurement) are known as "inter-coder reliability." The question is: would different coders, confronted with the same text, and using the same coding scheme, "tick off" the same coding categories? The answer to that question partly depends upon the design of the coding scheme and coding categories: the more abstract and theoretically defined the categories, the more likely it is that different coders will come up with different results. It is good practice to test the reliability of each coding category by having different coders code the same material.

An Example

Content analysis can be applied to a variety of texts (e.g. newspaper editorials, speeches, documents, letters, ethnographic field notes, transcripts from in-depth interviews or FOCUS GROUPS) or images (e.g. advertisements, photographs). Below is a transcript of a focus group of young British-born Pakistani women who married men from Pakistan. The participants are: Nadia (N), Laila (L), Ayesha (A), and the Moderator (M: Maria Zubair, to whom I am grateful for the transcripts).

M: How did your husbands find life here – when they got here?

L: I think it depends really on how they live … when they're in Pakistan. I mean some people, you know --- they're less fortunate than others, they live in villages or whatever, you know, they don't really go to college, but then there are those who live in towns and they have decent upbringing --- you know so when they come here they don't --- they are obviously – they're quite different but not so much. …

…

L: Yeah, I mean with my husband … when he came over here, he was the only child – um --- he'd been to college, university and everything so he's educated – so when he came here he didn't find it too much of a difference … So of course they find it different because it's … a lot different isn't it!

…

L: You know going out – they find it difficult to make friends, I think that's the main thing.

A: Yeah.

L: To make friends – or places to go – you know sometimes in the evenings and weekends we say oh where shall we go, but there's nowhere to go you know – clubs and pubs they're for English people they're not for us.

…

L: They find it difficult anyway because for them to come and mix with – um British born Asians is --- It's a lot difficult --- It is more difficult because they're just not on the same wavelength, you know, I mean my husband – he thinks Pakistanwise and the boys from here they think differently – don't they!

A: I find that – I find that with my husband as well

L: Yeah

A: Sometimes I just sit there and um --- we don't (laughs) – we're not on the same wavelength … he sees things the Pakistani way … The time he comes home, he just sits there in front of it [TV] – on sofa – right there. He won't move then. He just sits and that's it, you know. (laughs)

…

L: My husband used to say to me when he came over here 'oh I don't like it' … 'oh I don't like you wearing English clothes and stuff - oh why don't you wear *hijab* [head scarf]', and you know, all that stuff. 'Why can't you wear Asian clothes to work'. … now he never says that, you know (N: yeah). So he's adapted to it … but when they first come here they do get a cultural shock – I mean everything is different (M: yeah) from there, you know, so --- Over there it is, isn't it. It's totally different.

…

A: [(laughs) My husband - he, you know, he just winds me up. He used to wear *shalwar qameez* [Pakistani clothes] when he first got here – and I said to him, look when you go out … you're not going in these.

Even a simple word count reveals the dichotomous nature of the social context (perhaps not unsurprisingly given the Moderator's leading question). The words occurring most frequently are: "they" (21), "he" (16), "I" (12), "here" (9), "there" (6). The relative high frequency of such words as "different" (5) and "difficult" (5) also goes a long way in highlighting the complex social relations of a marriage between a British-born Pakistani woman and a Pakistani man. After all, such words as "same" and "easy" never appear in the text. Thematic coding categories would also help to tease out the peculiarities of these different worlds. Whether done with the help of a specialized computer software or simply on paper, a coding scheme based on thematic categories would have to tap the following spheres of social life: "family", "friendships", "jobs", "leisure". But these categories are too broad. The coding scheme would have to zoom in closer into the "Difficulty/Easiness" in "Marital relationships", "Own family and in-laws", "male and female friends", "friends in Pakistan and in England". Frequency distributions of these categories would offer further insights on the question: "How did your husbands find life here?"

Yet there is more to the meaning of a text. Take Laila's words: "some people … less fortunate than others." The choice of words is not random. Consider the alternative: "less brave", "less driven", "less adventurous". No: they are simply "less fortunate". There is no blame on them. Just the luck of the draw. Conservative meritocratic views of the world would have us believe that we are all dealt a fair deck of cards. What we do with them is up to us, to the brave, the daring, the entrepreneurial. Implicitly, Laila frames her discourse along more liberal, rather than conservative lines. But she also frames her discourse on the basis of implicit social-scientific models that we can easily recognize as such. People who grow up in villages don’t go to college. People in town have a decent upbringing. They get an education. They are not so different. Social scientific models of urbanization and of the relation between education and tolerance lurk behind Laila’s words. Finally, Laila frames her discourse in very dichotomous terms of “us vs. them”: “clubs and pubs they’re for English people they’re not for us”; “he thinks Pakistaniwise”; “we are not on the same wavelength”. Laila, and the other women in the group, and their husbands are different from the English. But they are also different from their husbands. And Laila’s educated husband is different from some of the other husbands …

These different modes of argumentation and framing could be highlighted through such coding categories as: “Conservative political views”, “Liberal political views”; “Inclusive worldviews” and “Exclusive/Dichotomous worldviews”; “Mode of reasoning (or argumentation)” with such sub-categories as “Emotive”, “Social scientific”. An emphasis on rhetoric and modes of argumentation would lead us to include such categories as “Appeal to religion”, “Appeal to science”, “Appeal to emotion”, “Appeal to reason”, “Appeal to values of equality”, …

Referential analysis could be used to tap the differential portrayal of social actors and situations. For each main actor (e.g. the English, the husbands, the wives, the friends, the relatives back in Pakistan), event (e.g. visiting Pakistan, getting married, moving to England), or situations (e.g. an evening at home, time spent on weekends, dressing and eating habits) we could have three basic coding referential categories: positive, negative, or neutral.

Conclusions

Content Analysis, as a typically quantitative approach to the study of texts, offers invaluable tools for teasing out meaning from texts or any other symbolic material. If confronted with the analysis of hundreds of pages of transcripts of the kind illustrated above (or of any other types of documents/symbolic material) the variety of content analysis techniques would certainly help us reveal patterns in the data. One advantage of quantitative analysis is that it is efficient: it helps to deal with large quantities of data without getting buried under the sheer volume of material. Its disadvantage is that it may miss out on subtle nuances in the production of meaning. Even the few examples discussed above should have made that very clear. To think in terms of a quantitative versus a qualitative approach to texts is a misguided approach. Each has strengths and weaknesses (consider this: how long would it take to do a qualitative analysis of hundreds of pages of transcripts? how would you ensure that you are not relying on the best sound bytes, or that you have not forgotten anything?). Theoretical/methodological developments in the analysis of texts coupled with technological developments in computer software are increasingly blurring the lines between quantitative and qualitative approaches (e.g. CAQDAS, Computer-Aided

Qualitative Data Analysis Software; Kelle, 1995; Fielding and Lee, 1998). At the current state of the art the best approach is probably one that combines a broad picture of a social phenomenon through quantitative content analysis with sporadic and deep incursions into selected texts using more qualitative approaches.

References

Abell, Peter. 1987. *The Syntax of Social Life. The Theory and Method of Comparative Narratives*. Oxford: Clarendon Press.

Berelson, B. "Content Analysis". Pp. 488-522 in G. Lindzey (ed.) *Handbook of Social Psychology*, Vol. 1. Reading, MA: Addison-Wesley.

Fielding, Nigel and Raymond Lee. 1998. *Computer-Analysis and Qualitative Research*. London: Sage.

Franzosi, Roberto. In press. *From Words to Numbers*. Cambridge: Cambridge University Press.

Holsti, O.R. 1969. *Content Analysis for the Social Sciences and Humanities*. Reading, MA: Addison-Wesley Publishing Company.

Kaplan, Abraham. 1943. "Content Analysis and the Theory of Signs." *Philosophy of Science.*Vol. 10, pp. 230–47.

Kelle, Udo (ed.). 1995. *Computer-Aided Qualitative Data Analysis*. London: Sage.

Krippendorf, Karl. 1980. *Content Analysis: An Introduction to its Methodology*. New York: Sage.

Shapiro, Gilbert and John Markoff. 1998. *Revolutionary Demands: A Content Analysis of the Cahier de Doléances of 1789*. Stanford, CA: Stanford University Press.

Weber, Robert Philip. 1990. *Basic Content Analysis*. Newbury Park: Sage.

Berelson, Bernard. 1952. *Content Analysis in Communication Research*. Glencoe, IL: The Free Press.

Budd, Richard W., Robert K. Thorp, and Lewis Donohew. 1967. *Content Analysis of Communications*. New York: Macmillan.

Franzosi, Roberto. 2004. "Content Analysis." In: pp. 547–66, Alan Bryman and Melissa Hardy (eds.), *Handbook of Data Analysis*. Beverly Hills, CA: Sage.

—. 2008. *Content Analysis (Benchmarks in Social Research Methods series)*. 4 vols. Thousand Oaks, CA: Sage.

Gottschalk, Louis August and Goldine C. Gleser. 1969. *The Measurement of Psychological States through the Content Analysis of Verbal Behavior*. Berkeley: University of California Press.

Gottschalk, Louis August, Fernando Lolas, and Linda Louise Viney. 1986. *Content Analysis of Verbal Behavior: Significance in Clinical Medicine and Psychiatry*. Berlin: Springer-Verlag.

Holsti, Ole. 1969. *Content Analysis for the Social Sciences and Humanities*. Reading, MA: Addison Wesley.

Krippendorff, Klaus. 1980. *Content Analysis. An Introduction to its Methodology*. Beverly Hills, CA: Sage.

—. 2004. *Content Analysis. An Introduction to its Methodology*. Second Edition. Thousand Oaks, CA: Sage.

Lasswell, Harold D. 1926. *Propaganda Technique in the World War*. The University of Chicago: PhD Dissertation.

—. 1927. "The Theory of Political Propaganda." *The American Political Science Review*, Vol. 21, No. 3, pp. 627-631.

—. 1930. *Psychopathology and Politics*. Chicago: The University of Chicago Press.

—. 1935. "Verbal References and Physiological Changes during the Psychoanalytic Interview: A Preliminary Communication." *Psychoanalytic Review* Vol. 22, No. 1, pp. 10-24.

—. 1936. *Politics: Who Gets What, When, How*. New York: McGraw-Hill.

—. 1936a. "Certain Prognostic Changes During Trial (Psychoanalytic) Interviews." *Psychoanalytic Review* Vol. 23, No. 3, pp. 241-247.

—. 1938. "A Provisional Classification of Symbol Data." *Psychiatry*, No. 1 pp. pp. 197-204.

—. 1948. "Structure and Function of Communication in Society", In: pp. 37–51, Lyman Bryson (ed.), *The Communication of Ideas: A Series of Addresses*. New York: Institute for Religious and Social Studies.

—. 1949. "Why Be Quantitative?" In: pp. 40-52, Harold D. Lasswell, Nathan Leites and Associates. *Language of Politics: Studies in Quantitative Semantics*. New York: George W. Stewart.

—. 1963. *The Future of Political Science*. New York: Atherton Press.

Lasswell, Harold D. and Dorothy Blumenstock. 1939. *World Revolutionary Propaganda: A Chicago Study*. New York: Alfred A. Knopf.

Lasswell, Harold D., Ralph D Casey, and Bruce Lannes Smith. 1935. *Propaganda and Promotional Activities: An Annotated Bibliography*. Chicago: University of Chicago Press.

Lasswell, Harold D., Daniel Lerner, Ithiel de Sola Pool. 1952. *The Comparative Study of Symbols: An Introduction*. Stanford: Stanford University Press.

Neuendorf, Kimberly A. 2002. *The Content Analysis Guidebook*. Newbury Park, CA: Sage.

Weber, Robert Philip. 1990. *Basic Content Analysis*. Newbury Park, CA: Sage.