# Computer aided content analysis of historical and process-produced data: methodological and technical aspects

Mochmann, Ekkehard

Ekkehard Mochmann

# Computer Aided Content Analysis of Historical and Process-Produced Data: Methodological and Technical Aspects

Document analysis as an intellectual as well as a computer aided procedure has a sufficiently long tradition to be recognized as a standard instrument in social research. Like interview and observation, content analysis is used to generate data for the analysis of social reality. Records from almost any source of communication can provide an empirical base for statistical analysis and subsequent interpretation. Content analysis, in particular the computer aided procedures, will be reviewed here as a possible candidate for inclusion in the set of instruments that will be needed. For this purpose an attempt is made to answer three questions:

1) What can quantitative history expect from content analysis?
2) What computer aided procedures are available?
3) How can content analysis be applied to historical and process produced data?

Suggestions for introducing this method into the research process dealing with historical and process produced data will be made. The final decision on whether and how it can be used will depend, of course, on the objectives of substantive research.

## 1. What Can Quantitative History Expect From Content Analysis?

Content analysis may well be the appropriate instrument for Social Science History whenever „textual data" not containing quantitative information per se (like statistical tables, turnover figures etc.) are under investigation. Text books on social science methodology recommend content analysis as the obvious method for research on communication. These can be communications from manifold sources

[1] Definitions of content analysis are discussed in e. g.: Markoff, J., et al., Toward the Integration of Content Analysis and General Methodology, in: Heise, D. R. (ed.), Sociological Methodology 1975, San Francisco 1974, pp. 4–7, and Holsti, O. R., Content Analysis for the Social Sciences and Humanities, Reading 1969, p. 2 ff.

like speeches, pictures, movies and other manifestations of symbolic behavior. In this contribution we will confine ourselves to written documents.

Without discussing the many definitions of content analysis we propose to label all procedures content analysis that comprise the systematic description, reduction and inspection of communication under the analytic frame of research concepts[1]. Subsequent analysis of the relations between these concepts may include inferences on origin, context and reception of the communication.

Part one of this procedural definition covers the aspect which attributes the distinctive feature to content analysis as compared to other instruments. As the process of interviewing generates data from responses, content analysis does the same when applied to texts. Frequently content analysis projects stop after displaying the frequency distributions for these data. It is obvious that it would be inappropriate to consider research finished after textual data is converted to numerical form. So far it has achieved not much more than information reduction. Analysis and interpretation have to start thereafter. Looking at content analysis in this way may present a subtle misunderstanding of the value of frequency counts on concept or word occurrences. They are as much a final result of research with content analysis, as the marginal distributions are the final result of research employing interviews. It is a question of research design of how much you get out of it by further analysis.

Before addressing this area, however, we shall recall in what ways reduction of text can be achieved.

It has been postulated that coding text by content analysis procedures should be semantically as close as possible to the contents of the original documents[2]. This basically means taking redundancy out of the text and boiling it down to its meaning constituents. If, in addition a standard descriptor language is used, heterogeneous information from the original representation becomes comparable. In this way, similar cases can be grouped together and thus can be counted. The intention of this approach is not to read between the lines or to involve analytical processes in the information reduction step. It is rather a condensation of verbose description into statistics as a result of a coding and counting process. Analytical and inferential steps may be based on these results, no longer on the original information. The question must then be asked whether after this kind of reduction instrumental content analysis is still possible. For representational approaches it may be the appropriate procedure. But content analysis can go beyond condensed semantic representation. It can be analytic in the reduction process. This is usually the case in traditional content analysis. Psychological applications for instance, have shown that it is possible to capture the pragmatic aspects when the coding process builds on the connotative meaning of words or the latent meaning of larger communication units.

Traditional content analysis, employing human intellect for the reduction process, requires the translation of the text under investigation into a predefined cate-

[2] Markoff, Integration, p. 3.

gory scheme. The categories representing the research variables are defined intentionally. A coding rule describes which set of denotation units from the text should be grouped under one category. The more familiar the human coder is with both the text under investigation and the classification scheme, the better the expected result. This is optimally the case, if a native speaker, familiar with the subject area, is conversant with the classification scheme as well.

But what performance can be expected from a coder who is not familiar with the context of a given text? He will hardly manage to understand the text entirely unless it is explicit and simple. If the text contains insider information, he will be lost on these grounds, even though he may have above average linguistic competence. He may not be able to get the latent or even the manifest meaning. Nevertheless he may well succeed in breaking down the text into smaller communication units and assign categories to them. This may be a problem of particular importance in the attempt to analyze historical documents when context information is lacking, whereas process produced data as a rule have, by administrative law, a clearly defined context.

Computer aided procedures are not able to handle complete texts as entities nor are they able to correctly identify the boundaries of meaning units aside from syntactical boundaries. We consciously exclude experiments in artificial intelligence which achieved some practical results, however, in a small, well defined domain. Thus in most applications the text is broken down into the single words by the input program. The subsequent programs then relate the denotative and connotative meaning of the individual words from the text to the categories defined a priori in a dictionary or they ascertain statistical associations and measures in an empirically exploratory way.

## 2. Which Computer Aided Procedures Are Available?

As suggested above, we may distinguish the following procedural steps of a content analysis: 1. Description, 2. Reduction, 3. Inspection. Along these lines, we will describe what options are available in the various software systems and what they can be used for.

### *2.1 Descriptive Procedures*

Automatic procedures process text as characters or character strings. Units which are separated by empty spaces or other delimiters are recognized as words. Thus most systems operate on the single word as the unit of information. The listing of

all words occurring in the text in different forms (token) as well as frequencies of identical word forms used (types) and alphabetically sorted indexes can be generated easily. Most programs generally available allow modifications of these lists according to ascending or descending frequency and forward or backward alphabetization. These frequencies allow the computation of various quotients like Type Token Ratio (TTR). The TTR is used for the characterization of the differentiated word usage or the restricted word spectrum of the text source. Drawing on extralinguistic information, e. g. stress at time of the generation of the message, the TTR can be used as a measure of stress intensity by relating the actual figures against the individual standard and computing the deviation. Various TTR computations for the same text source have proved to be relatively constant. Like Markov transition probabilities in the sequencing of words or arguments, they can be used for authorship identification. To support disambiguation and dictionary construction, Key Word in Context routines (KWIC) can be used. They list the words after permutation in their textual environment. KWIC, as well as KWOC (Key Word out of context) are available at almost every installation that offers text processing facilities[3].

### *2.2. Information Reduction by Categorization*

Empirical social research applies particular measurement techniques and categorizations to describe social reality. Interview surveys are used to collect data describing properties of individuals. The goal of measurement is the grouping of comparable characteristics in answer categories of variables, which then can be statistically analyzed to identify interdependencies between the variables. Applied to content analysis procedures, this categorization process can be characterized as the grouping of word occurrences (token, individual characteristics of the particular text) under stem forms (types, characteristics as combination of different forms), which can then be assigned to theoretical categories on a higher level (variables). Alternative procedures used for categorization will be exemplified by drawing on the routines in the most prominent system, the GENERAL INQUIRER[4].

[3] At most universities the department for Linguistics will have suited software. For special developments see also: Genet, J. G., Medieval History and the Computer in France, in: QUANTUM Information, 5 (1978), pp. 3–10.

[4] Stone, P. J., et al., The General Inquirer: A Computer Approach to Content Analysis, Cambridge/Mass. 1966.

### 2.2.1. A Priory Categorization by Dictionaries

The GENERAL INQUIRER system assigns text entries to theoretical categories by various program steps. The relation between the possibly relevant entry words and the categories is predefined in a content analysis dictionary. About 30 different dictionaries are in existence[5]. The more recent dictionaries contain context information for disambiguation of homographs[6].

Not all words occurring in the text base can be anticipated when creating the dictionary. These ‚new words' will be displayed in a leftover list. On this basis the investigator can make few tag-entry assignments and incorporate them into the dictionary. Results of the tagging operation are displayed in the TAG TALLY LIST. It gives frequencies in absolute figures and percentages on the basis of the total number of words and sentences in the text. The results are stored on magnetic tape or disk for post-tagging operations, such as retrieval of particularly interesting text parts for closer inspection, further analysis of purposely selected subsamples, and interfacing to statistical analysis packages.

The principles of the dictionary approach developed for the GENERAL INQUIRER were adopted by systems like EVA, TEXT, TEXTPACK and SPENCE's[7]. While the idea of the GENERAL INQUIRER was to offer a general content analysis instrument, the more recent developments were initiated for special application needs. EVA was developed for the analysis of newspaper headlines, ANACONDA and TEXTPACK were developed for the coding of answers to open ended questions, and SPENCE's programs for the analysis of psychiatric interview protocols. Since they are special purpose oriented, they developed certain special features further while neglecting others that are needed for more general applications.

TEXTPACK offers a very efficient set of routines for dictionary comparisons, text correcting and selection of particular answer texts. EVA was aiming at further developments for an advanced semantic analysis of headlines. Both systems offer direct interfaces to statistical analysis packages (EVA – RAPROSYS, TEXTPACK – OSIRIS). The SPENCE programs are particularly suited for smaller texts (up to 100 lines per segment). Even though COCOA, written for linguistic analysis, does not offer tagging routines it is appealing to use for teaching basic procedures in content analysis[8]. It offers a very flexible parameter language for structuring and labelling the input as well as all sorting and index list options.

[5] An overview of General Inquirer Dictionaries (status: Fall 1965) is given in Stone, op. cit., pp. 140–141.

[6] A detailed description is given in: Kelly, E., and Stone, P.J., Computer Recognition of English Word Senses, Amsterdam 1975.

[7] Methods and techniques of available content analysis software are reviewed in: Mochmann, E., Automatisierte Inhaltsanalyse, in: Langenheder, W. (ed.), SIZSOZ Expertisen: Ausgewählte Gebiete sozialwissenschaftlicher DV-Anwendung, Vol. 1, St. Augustin 1976, pp. 61–92.

[8] Berry-Rogghe, G. L. M., and Crawford, T. D., COCOA Manual, Chilton, Didcot, Berkshire. A tagging routine has been added in Cologne.

### 2.2.2. Inductive Categorization by Statistical Procedures

Parallel to dictionary systems, the empirical inductive procedures were developed. They explicitly avoid an a priori categorization[9]. The criticism of dictionary approaches can be subsumed under the following arguments: A priori dictionaries are derived from or oriented to sociological or psychological theories. Thus they reflect particular research intentions. Empirical approaches are neutral in that respect. In addition there are serious doubts whether dictionaries can appropriately anticipate the vocabulary domain of the sociolinguistic community that produced the text under investigation.

The leading, and so far only inductive system, is Iker's WORDS system[10]. The conceptual design of WORDS was based on the intention to exclude any influence of the researcher on the derivation of concepts from the text. The themes and theoretical concepts should be derived from the texts by means of statistical procedures. The unit of information is the single word, based on the assumption that sufficient meaning resides in the words and in the associations among and between words to produce an accurate representation of content. The procedure can roughly be characterized by the following steps:

1) An input document is divided into „segments", e. g. the page, the paragraph, equal numbers of words, time units in order to achieve comparable length of documents;
2) all function words, e. g. articles, conjunctions, are removed;
3) remaining words are deinflected to root form (lemmatized);
4) the frequency of occurrence in the segment is computed for each different word;
5) a subset of these words (types) is selected for analysis;
6) an intercorrelation matrix (ICM) is computed on this selected subset;
7) the ICM is then factor analyzed (principal components algorithm) and
8) rotated to simple structure against a varimax criterion.

The resulting factors have been shown to correspond with major content themes in the document under analysis. Cluster algorithms have been applied successfully as well.

Whereas the empirical approach tends to have advantages over the dictionary approaches with respect to neutrality, the problems are here how to interpret the results of the factor or cluster analysis. Furthermore, problems arise from the fact that any empirical procedure is plagued by how many variables to incorporate. Certain approaches become inapplicable if the number of variables exceeds the number of cases. In content analysis, the number of documents or sections of documents is likely to be much smaller than the number of different words. Thus some selection

[9] Iker, H. P., Harway, N. J., A Computer Systems Approach Toward the Recognition and Analysis of Content, in: Gerbner, G., et al., The Analysis of Communication Content, New York 1969.

[10] Iker, op. cit.

is sensible. Selection can easily be oriented towards topical words, stylistic words, or whatever, but selection becomes a crucial determinant of the resulting description.

Iker has proposed some interesting empirical criteria for selecting words that enter into an empirical procedure. He correlates all terms with each other and then sums the 5th power of the correlation for each term, selecting these terms that have the largest sums. He selects the terms that have a subset of high sums of correlations rather than those terms that have many small inter-term correlations (SELECT procedure).

### 2.2.3. A Contextual Approach to Content Analysis

All methods discussed so far operate on the unit word. Contextual information is incorporated only modestly. This is where Cleveland, McTavish and Pirro come in with their QUESTER system[11]. Since the communication content is permanently changing, a communication model that attempts to analyze the communication process should be dynamic as well and should pay attention to the context. Words in a dictionary do not have a natural context, whereas words in communication do (exception: the HARVARD IV DICTIONARY contains disambiguation context). In addition words in a conversation do have known properties like syntactical structure, conceptual structure and contextual structure. These structures are interdependent, and like certain syntactical structures define the structures following (e. g. S-P-O), certain concepts are activated by preceding concepts. According to Quillian, a model should define which concepts are called by the nets of words in the context of a particular word. The relevant context defines these nets. QUESTER uses distance measures to define net boundaries.

### 2.2.4. Coding Approaches

Particularly for historical applications Couturier and Abehassera have developed the FORCOD system. Trained coders report in a standardized terminology on a tape recorder while reading the source text. The contents are represented in pairs of ‚definers' and ‚descriptors' (e. g. definer: occupation, possible descriptors: merchant, wine-merchant, wine-grower, carpenter). Then codes are assigned arbitrarily by the program to these descriptors. They may be recoded by intervention of the researcher to meaningful codes. These codes are analysed by the table analysis program FORTAB.

[11] Cleveland, E., et al., QUESTER: Contextual Analysis Methodology. Paper read at the ISSC Workshop on Content Analysis in the Social Sciences, Pisa 1974.

### 2.2.5. Retrieval according to subject for closer inspection

Over the last ten years powerful retrieval systems have been developed, like GOLEM and STAIRS, which are products of hardware manufacturers, RIQS, TEXT and Z.A.R., which have been specifically developed in the social science community[12]. According to a request by subject, they identify and retrieve from a document pool those documents which address the same topic. This function is available in some content analysis systems too (e. g. GENERAL INQUIRER). They may be misused to count how many documents were addressing a particular topic.

In particular, since many administrations have started to store information in computer accessible information pools these systems gain increasing importance. Prohibitively high costs may prevent their application for content analysis purposes on a larger scale even though the necessary routines are available in their software.

## 3. How Can Content Analysis Be Applied to Historical and Process-Produced Data?

The recent QUANTUM Documentation on Quantitative History 1977 lists several projects employing content analysis[13], among these:

- Social structure of NSDAP. Analysis of its elites (Kalusche).
- Analysis of Prussian school books under the aspects ‚education and industrialisation' (Lundgreen).
- Quantitative analysis of SA-elites social structure (Jamin).
- Abitur 1917–1971. Content analysis of graduation compositions (Mohler).
- Resistance in National Socialist Germany (Mann).
- Interest conflicts in trade politics during the German Revolution in 1848/49 (Best).
- Social status of candidates for the Reichstag 1898–1912 (Schroeder).
- Text analysis of Middle Latin chronicles (Arnold).
- Social Protest in 19th century Germany (Tilly).
- Marriage and family in German bishops letters to their flocks (Schmaelzle).
- The rise of a new elite: Social structure and political function of provincial administrators in France 1787–1820 (Reichardt).
- Content analysis of wills from 1648 to 1791 (Vogler).

[12] These systems operate on full text natural language. The retrieval process can be viewed as a reversed indexing (coding) process. The number of documents retrieved is displayed for each retrieval query. This could be used for content analysis purposes.

[13] Bick, W., et al., Quantitative historische Forschung 1977, Stuttgart 1977.

Those projects can be distinguished according to two goals. Most projects are analyzing phenomena for which statistical information is lacking. So they had to derive it from descriptive sources by a coding process. A minor number of projects is concerned with the orientation of the text source itself. For the time being the coding approach seems to be prevailing. This may be explained by the objective of revealing the social structure of groups primarily involved in historical events. The analysis of values or underlying intentions of the communicator has not yet been often focused upon[14].

Some of the historical documents referenced in the QUANTUM documentation date back to the 12th century. The average time range for the period under investigation was calculated to be 114 years[15]. These time ranges put even more emphasis on the requirement to carefully consider relevant contexts. For historical documents the following should be considered as particularly relevant contexts[16].

| | | |
|---|---|---|
| | | Time (situational) context |
| | Implicit context | |
| | | Space (physical) environment |
| Total context | | |
| | | Linguistic (verbal) context |
| | Explicit context | |
| | | Extralinguistic context (kinesics) |

Since the computer programs are designed to analyze the linguistic context, the researcher himself has to control the impact of the other contexts. Given that natural languages are a dynamically developing code (Latin may have been the only exception in the Middle Ages) we have to pay attention to changes in vocabulary and shifts in meaning over time. When investigating more recent documents, the researcher will be aware of socially redefined connotations. The latest significant example in the Federal Republic of Germany is the mention of ‚Kreuth'. Before the decisive meeting of the CSU it was just a name of a village – at most important from a tourist's point of view. Thereafter it became synonymous with a dramatic discussion about the split between CDU and CSU.

We have to control these effects in order to avoid possible serious errors. This may imply thorough analysis of major events in a given period of time in addition to the documents under investigation. On the other hand we may draw inferences on unexpected changes of significance in word associations. Maybe they are clues to unknown social processes in the past for which no empirical evidence is yet available.

[14] Namenwirth, J. Z., Some Long- and Short-term Trends in One American Political Value: A Computer Analysis of Concern with Wealth in 62 Party Platforms, in: Gerbner, Analysis, pp. 223–241.

[15] Bick, Forschung, p. 12.

[16] Cf. Slama-Cazacu, T. S., Die dynamisch-kontextuelle Methode in der Sprachsoziologie, in: Kjolseth, R., and Sack, F., Zur Soziologie der Sprache, KZfSS, Special Issue 15 (1971), p. 82.