Observe!
A Popperian Critique of Automatic Content Analysis

Peter Ph. Mohler & Cornelia Zuell

ZUMA, Postfach 12 21 55, D-68072 Mannheim

Abstract

There is a revival of automatic content analysis, at least for open ended questions. Automatic text analysis intends to analyse texts (written material) "untouched by human hands". It is an alternative to the dictionary approach, where categories intend to measure predefined concepts (computer assisted content analysis). The basic procedures of some prominent programmes for automatic content analysis will be explained and discussed. Both, automatic and computer assisted content analysis will be compared. The data base is a special issue of the Behavioral American Scientist on "exporting social surveys" (American Behavioral Scientist, Vol. 42, No 2, Oktober 1998, Sage Publication Inc.)

Keywords: content analysis, dictionary-based, automatic

1. Introduction

In his critique of observation without theory Popper (1972) tells the reader to "observe". There is an immediate reflex to this request: "what should be observed?", followed by others like "why", "how" and "when". To answer Popper's simple request just to observe, one obviously needs quite some sophisticated ideas about the object under observation and the tools to observe it. In his discussion of the phenomenon Popper point to the "inbuilt" theories in our brains, eyes and other means of observation.

Today, we observe a renaissance of "theory-free" content analysis tools. Real 'automatons' which require nothing but textual input. All the analysis is done "untouched by human hands" as Iker and Harway called this kind of text analysis (Iker et Harway, 1969).

The general approach of such an automatic content analysis within the field of computer assisted content analysis in the social sciences competes with the "dictionary approach" as proposed by Stone and many others since the early Sixties (Stone et al., 1966). The dictionary approach insists on theory driven development of categories which serve as classification tools (Mochmann, 1980).

However, if observation without theory is impossible or at least non-popperian (i.e. not part of the main stream concept of science), questions raise like what are the "inbuilt" or hidden theories of the automatic tools on the market today? What is their analytical power? Are they limited to thematic analyses, as Iker and others indicated? What properties must a text have to be suitable for automatic analyses?

2. Approach

We will answer these questions in a twofold way. Firstly, we will uncover and discuss the inbuilt procedures, hypotheses and theories of Iker's Standard Approach from 1969.

Secondly, we will apply his automatic approach to a small topical well defined text corpus and contrast it with a dictionary approach.

The textual data base is one issue of the American Behavioral Scientist (No. 42) on "Exporting Surveys"[1]. We took this as an example of a coherent, but not very strict, assembly of texts dealing with a general topic. It is long enough to allow for statistical analyses three levels (article, paragraph and sentence). It is, on the other hand, short enough to validate the results intellectually. Two different analyses will be presented. Firstly, an "Untouched by Human Hands approach" (as described by Iker et Harway, 1969) and secondly a dictionary analysis. Iker's approach is taken here, because it is the least restrictive among the automatic approaches.

3. Examples of programmes for automatic content analysis

Most of the tools available today are limited to a special type of text (e.g. answers to open-ended questions) or to specific, predefined categories. Here we will give some examples how some of these programs work (see also Alexa et Zuell 1999).

The **Words** program of Iker et Harway (1969) is the only one which makes an exception from the restriction mentioned above. Its technology aims at all kinds of structured text with a certain amount of redundancy (i.e. repetitiveness). While the programme itself is no longer marketed, one can simulate it rather easily. The basic procedure is as follows: all input text has to be divided into coherent text segments. Iker himself took different window sizes reflecting the structure of a text (Iker, 1974). One then eliminates words *non functional* for a thematic analysis (in Ikers ideology articles, punctuation, etc.). A frequency list of words is produced. Each word forms its own category - however, strict synonyms are lumped together in mini-categories. The n most frequent mini-word categories are then identified in each text segment. The result is a data matrix of text segments as cases and mini-word categories as variables. An intercorrelation matrix is computed for the co-occurrence of all the variables per text segment. The resulting matrix cannot be input directly into multi-variate statistical procedures because of the many weak correlations between variables. Such weak correlations indicate "ever present" and thus not differentiating words (like the word 'Kant' in a biography of the famous philosopher). They function as a kind of noise or smog screen obscuring the vision on the really "strong" correlations. Iker used, after some experimenting, a procedure, where each correlation coefficient was set to the power of five. By doing this, the small correlation coefficients, say .2 or .3, will become very small, while high correlation coefficients decrease not so much in their value. In the next step, all coefficients of one variable (i.e. mini-word category) are summed up. The variables are then sorted according to this sum. The n variables with the highest sum scores are then input into multi-variate statistics. The results can be interpreted as themes relevant or irrelevant for specific text segments.

Another program offering automated procedures, **TextSmart,** has been developed by SPSS Inc. It is limited to answers to open-ended questions and uses linguistics technology, such as word stemming, and statistical algorithms, e.g. clustering and multidimensional scaling, for generating automatically 'categories' for the coding of survey responses. TextSmart is advertised as "dictionary-free", in the sense that there is no need to create a coding scheme or 'concept dictionary' before running the analysis. A clustering procedure produces categories based on word co-occurrences. This is a three-step process: The program creates a matrix of

[1] American Behavioral Scientist, Vol. 42, No 2, Oktober 1998 published by Sage Publication Inc.

similarities from the terms (words, aliases, and stems) in the included terms list. It pairs each term with every other term in the list and checks to see how often each pair occurs in a text (a response), that is, how often it co-occurs. It constructs a contingency table for each pair of terms in turn. It uses this information to compute a binary measure, the Jaccard similarity measure (TextSmart User's Guide, p. 45ff), for each pair of terms. The measure consists of the number of co-occurrences between two terms divided by the sum of co-occurrences plus non-co-occurrences. In the second step, the program hierarchically clusters the similarity matrix and places the clusters into a user-specified maximum number of categories. The cluster algorithm used attempts to produce clusters whose largest distance between any two members is as small as possible; it tends to produce 'compact' clusters. In a third step TextSmart displays clusters using multidimensional scaling in two dimensions to scale the matrix of similarities. Nevertheless, the results of the automatic categorisation are usually not good enough and can serve only as a possible basis for developing 'meaningful' categories. One should consider that good excluded term and alias lists are of primary importance for the coding.

An third program of this type is **DICTION** which has been developed by Roderick Hart at the University of Texas. DICTION attempts to determine the language characteristics of texts on the basis of calculated scores for five general categories, namely Activity, Certainty, Commonality, Optimism, and Realism, comprising twenty-five sub-categories. The program incorporates dictionaries for each sub-category with a word list characteristic of each sub-category. The result of analysis with DICTION is a report about the specific text it has processed and a numeric file for statistical analysis. DICTION analyses the words of a single text of maximum 500 words based on the program's own dictionaries which are organised as word lists. There are twenty five such words lists, e.g., the Communication one contains words such as 'advice, recommend, urge'. The program matches the words in the lists with the words of the analysed text. The result of the matching of the words in the lists with the words of a particular text is the calculation of a number of scores for raw frequency counts and comparative data based on texts which have been previously examined by DICTION. On the basis of scores for the individual sub-categories, DICTION calculates a standardised score (normative data) for the five general categories.

4. Preliminary Results

4.1 Untouched by Human Hands Approach

Iker suggest not to use an oversimplifying automatic approach, were the association between words is not taken into account properly. He suggested to set the coefficients of association into the power of five and to compute the respective sums for each word. He argues that such a selection rule together with synonym/alias lists are needed to make the results intelligible.

To simulate Iker's approach we firstly decided paragraphs to be the observation windows. Then we produced a list of word frequencies without function words (like articles, conjunctions etc.). From this list we took the 60 most frequent words as 'mini word categories' (cutting at a frequency of 31). Synonyms were added, if applicable, to the mini word categories. Bivariate Pearson Product-Moment correlations were computed for all 60 categories per paragraph. Iker's selection rule mentioned above was subsequently applied to the 60 by 60 matrix of bivariate correlations. In a final step the top 24 categories were input into a MDS analysis (Graph 1).

One can easily identify a areas of connected word-categories (like EAST, WEST and DATA). That one author heavily relies on German examples for comparison, while others are not is indicated in the distance placing of GERMAN. However, it is not self evident that all the texts dealt with "exporting surveys", i.e. comparative survey methods. Prior knowledge, however, indicates that most articles don't deal with this issue (cf. Table 1 - headlines of articles analysed).

Graph 1: Iker approach

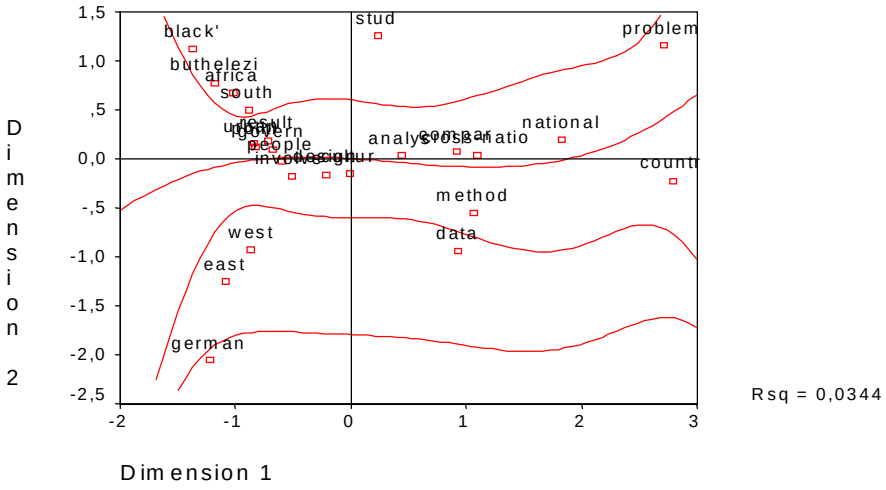


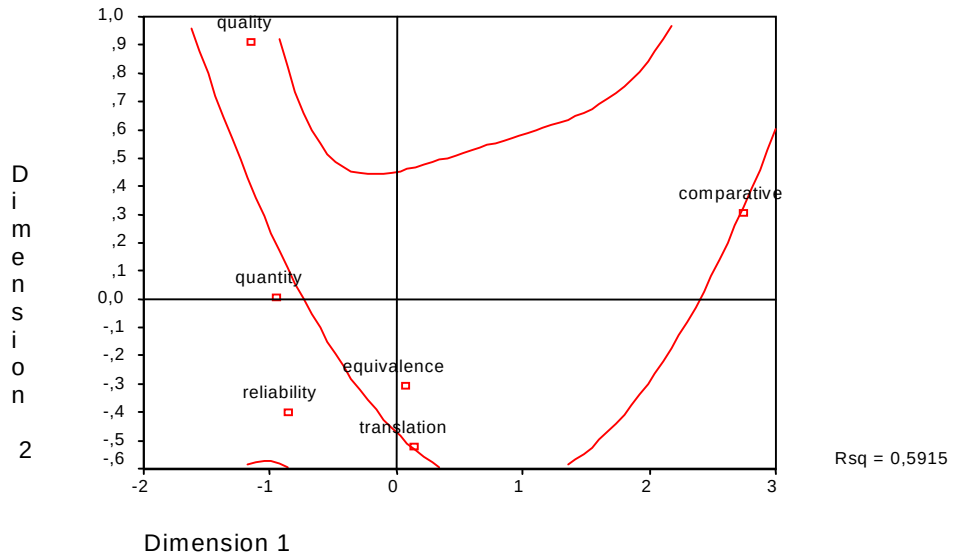
Table 1: Title of articles analysed

Buckley, Cynthia J. - Ideology, Methodology, and Context Social Science Surveys in the Russian Federation
Bulmer, Martin - Introduction: The Problem of Exporting Social Survey Research
Escobar, Roberts - Surveys in Mexico
Jowell, Roger - How Comparative Is Comparative Research?
Kuechler, Manfred - The Survey Method: An Indispensable Tool for Social Science Research Everywhere?
Newby, Margaret; Amin, Sajeda; Diamond, Ian; Naved Ruchira T. - Survey Experience Among Women in Bangladesh
Orkin, Mark - The Politics and Problematics of Survey Research: Political Attitude Studies During the Transition to Democracy in South Africa
Schooler, Carmi; Diakite, Chiaka; Vogel, Jerome; Mounkoro, Pierre; Caplan, Leslie - Conducting a Complex Sociological Survey in Rural Mali: Three Points of View

4.2 Dictionary Approach

A dictionary approach would take advantage of even more prior knowledge about general theme. The general theme is laid out in Bulmer's article on exporting social survey research. One could consult standard textbooks on comparative survey research, if available or read other literature on the topic. This would result in structuring the area along key-words or conceptual maps. For instance, one could be interested in the context of TRANSLATION as one of the crucial procedures in comparative research. Which are the other categories closely linked to it? In Graph 2 we present the MDS for a very small dictionary, almost nothing but a least of relevant key words of an abstract. Here TRANSLATION is linked to QUALITY and EQUIVALENCE, but not to QUANTITY and RELIABILITY. This reflects the known state of the art in comparative survey research. Translation is still be thought to be part of the "soft" procedures not being able to assess it quantitatively (which, by the way is actually not true).

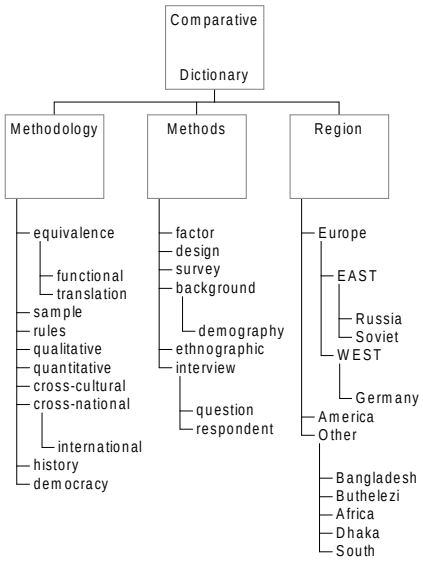
Graph 2: MDS Using Simple Key-Word Dictionary



However, this is a first step towards a conceptual dictionary approach. An a priori dictionary requires first a conceptual layout of what could/should be the case. In Popperian terms it is the explicit definition of what has to be observed, before any observation is made. Graph 3 represents such a conceptual layout for comparative survey research.

Three conceptual areas are mapped here: a. methodological issues, b. methods and techniques, and c: regions which are compared. Three types of analyses are offered by this approach beyond Iker's thematic analyses: a. Which of the pre-defined concepts are present in the articles, b. which of them are not dealt with, and c. which topic co-occur in one essay respectively across essays.

Graph 3: Dictionary conceptual map



We will discuss here the third type of analysis. In table 2 co-occurrences of dictionary categories are represented as correlations between categories and articles. It is evident, that despite the general theme, there is not much communality between the essays. Each of them deals with a different bundle of topics (as represented by the categories). Bulmer's introduction (article 4) itself deals with all topics, hence the correlations must be low.

Table 2: Dictionary Correlations between Categories and Articles

	Küchler	Jowell	Orkin	Bulmer	Buckley	Newby	Escobar	Schooler
Comparative		**.35**						
Qualitative						**.20**		
Survey			**-.15**					**-.10**
Interview						**.19**	**-.16**	
Analysis		**.15**						
History							**.11**	
Democracy	**.17**							
Cross-national		**.52**	**-.10**			**-.14**		
Sample						**.11**		
Cross-cultural					**.11**			**.10**
Equivalence		**.26**						
Europe	**.19**					**-.14**		
Others	**-.14**	**-.12**	**.23**		**-.16**		**.14**	

Pearson Correlation Coefficients. Basis 1342 paragraphs in 8 articles. Reported are here coefficients equal or higher than .10, level of statistical significance is .001

Three categories, BACKGROUND, RELIABILITY, and QUANTITY, do not correlate higher than .10 (which is already rather low) with any of the eight articles. From an a priori point of view, these are, however, rather important issues in comparative research. For instance, harmonising and standardisation of background variables (or demography) is a prerequisite of any comparative survey research. Not dealing at length and depth with it indicates a blind spot in the discussion presented by the authors. The identification of such missing topics cannot dealt with by automatic/Ikerian approaches in a systematic way. Because, systematic identification of what is missing, requires a theory about what should be not missing.

5. Conclusion

The conclusion is, that automatic content analysis might be helpful to identify areas of relevant themes in a text corpus. However, even then substantial results require quite some intellectual effort for revising synonym/alias lists or stop word lists. This effort should not be underestimated. On first view, automated approaches look very convenient on first glance. However, decisions have to be made about what are 'non-functional' words and, moreover, what is a 'synonom'. Both imply theoretical considerations a priori to the analyses, which go beyond the application or development of algorithms. Prior knowledge about the text data base is thus a prerequisite for both, the automatic and the dictionary approach. The automatic approach is a useful tool to identify topics and themes dealt with in a text data base.

In addition to thematic analyses computer assisted content analysis using dictionaries allows conceptual analyses of texts. Moreover, the dictionary approach provides a tools for identifying topics missing at all, because researchers define the possible universe of topics or concepts a priori. In a much more limited way this can be done with an automatic approach. One can identify, if one or a couple of texts do not deal with themes central to other texts in a data base.

As said above, both approaches require theoretical considerations. Thus 'untouched by human hands' doesn't of course mean no work for human brains. This also points to the well known fact, that the tools, be it a text analysis programme or a statistical procedure must fit the hypothesis and observation theories in the popperian sense to be able to observe.

References

Alexa, M. and Zuell, C. (1999). A review of software for text analysis. ZUMA.

Iker, H.P. (1974). SELECT: A Computer Program to Identify Associationally Rich Words for Content Analysis. *Computers and the Humanities*, Vol. 8:313-319.

Iker, H.P and Harway, N.I. (1969). A Computer Systems Approach Toward the Recognition and Analysis of Content. In Gerbner, G.A. et al. (eds.), The Analysis of Communication Content. Wiley & Sons.

Mochmann, E. (1980). Methoden und Techniken automatisierter Inhaltsanalyse. In Mochmann, E. (ed.), Computerstrategien für die Kommunikationsanalyse. Campus Verlag.

Popper, K. (1972). Objective Knowledge. In Popper, K.: Arthur Compton Lecture "Clouds and Clocks". Claredon Press.

Stone, P.J., Dunphy, D.C., Smith, M.S., and Ogilvie, D.M. (1966). The General Inquirer: A Computer Approach to Content Analysis. The M.I.T. Press.

TextSmart™ 1.0 (1997). User's guide. SPSS Inc.