Programs for content analysis available in the Netherlands

I.N. Gallhofer

This paper lists some programs available in the Netherlands for content analysis. Before introducing these programs we intend to characterize briefly the two main directions in computer aided content analysis i. e. programs based on dictionaries like the General Inquirer (1) and word procedures which do not make use of a priori categories like the program Words (2).

The dictionary procedure

This procedure will describe the General Inquirer system which is the most developed program in this field and exemplary of the type. Investigators using this technique focus on making inferences about messages they wish to identify in the documents. This specification is usually represented as a system of categories i.e. a dictionary where words are classified into theoretically relevant categories. Then the text words which are prepared following recording units and certain rules of format are assigned to category descriptors (tags) by a dictionary lookup procedure.

The meaning of words even though written in the same way (homographs) may be ambiguous and depends on the context. Therefore a sense distinction - or disambiguation procedure is built in the General Inquirer system before assigning tags to the text words. This is done by a set of sophisticated rules (syntactical and lexical ones) instructing the computer to examine the current context of the word and, based on what it finds, to make a decision regarding the intended sense (3). An other facility which contributes to processing efficiency consists of a suffix chopping routine. It removes the regular English inflections of words (i.e. -s, -ed, -ing, -ly, -er, -est). The stems of word roots are then looked up in the dictionary (4).

After chopping and disambiguating, tags are assigned to text words and frequencies of occurences and various kinds of statistical indices can be computed. The revised version of the General Inquirer program is available in Edinburgh for analyses of English documents. It provides a wide range of facilities: the input of a great variety of dictionaries, automatic disambiguation and chopping and even modification of the rules of those facilities (5).

The word procedure (6)

Iker and Harway developed the program Words in order to analyze texts without the use of a priori categories. This approach is based on the assumption that the meaning of words is sufficiently determined by their empirical associations to other words to allow the elicitation of major content themes and categories.
The procedure used is as follows: the input documents are divided into units of analysis. They may be segments of time of equal numbers of sentences etc. The recording unit is the single word. The operations of the program are as follows: First the program produces a data matrix which indicates the frequencies of occurence of the variables (words) and the several units (segments of text). Subsequently intercorrelations are computed which represent the degree of co-occurence between words as observed across successive units of the data base. Finally this matrix is factor analyzed in order to determine whether there are common factors which provide a meaningful interpretation of the text.
Since the program can only handle a matrix with 215 variables, before factor analysing, the number of words occuring in the text must be reduced. Program Words possesses some subroutines which remove all structural words like articles, prepositions and conjunctions. In turn another routine removes the inflections (chopping). Further reduction can be attained by applying synonymization in which generic words are created to subsume a set of semantically highly related high frequency words. When some meaningful factors are found which indicate themes or categories the researcher might be interested in the sequential appearance of the themes or categories in the data base. This can be realized by computing factor scores.
The program Words is available in Cologne. When one adapts the chopping and synonymization procedures analyses can be done with documents in different natural languages.
After having described the two main trends (7) in computer aided content analysis we can proceed to the evaluation of some available programs. What are the necessary prerequisites of a computer program for content analysis ?
In table 1 these requisites are summarized for the dictionary procedure and in table 2 for the word procedure.

Table 1 : Required capabilities of a program for the dictionary procedure

Required capabilities	comment
to match in an efficient way the dictionary words with the words in de documents	A dictionary mostly contains $\pm$ 5000 words; therefore an efficient procedure is of great importance.
alphabetically sorting of words	This routine contributes to the efficiency of the word matching procedure.
removal of inflections (chopping or stripping)	Chopping also has an impact on the efficiency of the program as one does not have to incorporate all the inflections in the dictionary.
disambiguation of word senses	It is of utmost importance to know in which context homographs are placed in order to determine the semantic category to which it belongs.
to search combinations of words in an efficient way	When one establishes as recording unit patterns which form a theme then this procedure is indispensable for an accurate frequency count.
the possibility to introduce different dictionaries	A researcher might develop his own set of theoretical relevant categories and he also might analyze documents in different languages.
to provide a good connection with other programs of data analysis like SPSS	If such a connection exists one can quite easily subject the data to further analyses.

Table 2 : Required facilities of a program for the word procedure

Required facilities	comment
removal of inflections	For the frequency count of words a chopping procedure is very useful because otherwise every different inflection of a word will constitute a variable.
synonymization	This facility also reduces the variables by subsuming generic words with the same meaning under one label.
efficient frequency count of $\pm$ 500 word roots	After removal of inflections and synonymization a frequency count of a smaller amount must be performed efficiently.
to provide a good connection with other programs of data analysis like SPSS	With such a facility a great variety of clustering techniques can be used.

Tables 3 and 4 indicate which requirements the available programs fulfill. From these tables one can conclude that the available programs hardly satisfy the requirements. We shall start with the discussion of Content.

Content

Content is a Fortran program for text analysis written by C.E. Cleveland and E. Pirro at the University of Minnesota. It is improved and revised by R. Marshall at Drake University and installed at the Sara computing centre in Amsterdam. Further information can be given by the author of this paper.

Content can handle dictionaries of maximally 250 categories with 8000 entries in total. For English texts the dictionary words can be entered in root form since the program provides a chopping procedure for the follo-

wing English inflections : -s, -es, -e, -ed, -ion, -ing. For Dutch and other natural language texts it still is necessary to enter inflectional forms (8). A serious defect consists in the absence of a disambiguation routine. Searches for patterns of words, different kinds of frequency counts, the computation of z scores and the printout of leftover lists of uncategorized words can be obtained. Since the removal of inflections only works for English words in connection with a dictionary lookup procedure and since there is no connection with other programs of data analysis Content only seems useful for the dictionary approach. A disambiguation could be performed - however cumbersome - with one of the other programs.

Table 3 : Comparison of the required facilities for the dictionary procedure with the facilities the available programs provide

	Content	Riqs	Cocoa
to match in an efficient way $\pm$ 5000 dictionary entries with the words in the documents	yes	no	no
alphabetical sorting of words	yes	yes	yes
removal of inflections	yes^{x}	no	no
disambiguation of word senses	no	yes^{x}	yes^{x}
to search efficiently combinations of words	yes^{x}	yes^{x}	yes^{x}
to provide the possibility to introduce different dictionaries	yes	no	no
to provide a good connection with other data analysis programs	no	yes	no

By x we indicate that this requirement is partially met.

Table 4 : Comparison of the requirements for the word procedure with the facilities the available programs provide

	Content	Riqs	Cocoa
removal of inflections	yes[x]	no	no
synonymization	no	yes[x]	no
efficient frequency count of ± 500 words	yes	yes	yes
a good connection with other programs of data analysis	no	yes	no

By x we indicate that this requirement is partially met.

Riqs

Riqs stands for Remote Information Query system. It is a retrieval program developed at Northwestern University at Vogelback Computing Centre under the supervision of Lorraine Borman. In the Netherlands it was introduced by dr. C. Middendorp at the Steinmetz data archive. It is now installed at the Sara computing Centre of Amsterdam. Information can be given by the Technical Centre of the University of Amsterdam.
As table 3 shows Riqs cannot handle efficiently dictionary entries. It also has no chopping facilities and the disambiguation requirement is only partially met. Since its output connects readily with SPSS and synonymization combined with frequency counts can be performed, it is more adequate for word analysis. An illustration of this procedure performed with Riqs is given by the author and W.E. Saris in Acta Politica no. 754, pp.491, 1975 and Z. Namenwirth's contribution in this issue.
Disambiguation can be done by using a key word in context index (KWIC).

This consists in an alphabetical listing of words of the documents. On the left and the right side of the key word appears the context illustrating each occurence of the word in the text. By means of this listing one can see if the specific usage of the word is in agreement with the category definition. If not there are two possibilities:
(a) this specific meaning of the word belongs to an other category or
(b) this meaning is irrelevant for the research in question.
If (a) occurs one might correct the specific word in the datafile and in the dictionary by adding a marker e.g. execute 1 and execute 2. In case of (b) it is best to delete this text segment.
The KWIC index also provides a listing of the different inflections of the words. In order not to enlarge unnecessarily the number of dictionary entries only the inflections for the specific texts to be analyzed could be incorporated in the dictionary. This would be a kind of substitute for chopping.

Cocoa

Cocoa is an acronym derived from a word count and concordance generator at Atlas. The revised CDC implemented version used at the Free University of Amsterdam is written in Fortran by Atlas Laboratory in Chilton (E. B. Fosey) and the University College at Cardiff (R.F. Churchhouse).
As no dictionary can be matched in an efficient way with the words in the documents this program only provides facilities for disambiguation by means of a key word in context listing called concordance.
Concerning word procedure Riqs is preferable to Cocoa since the latter's synonymization can be done by hand only (it is impossible to create new variables by a search procedure) and Cocoa output cannot be input into another data analysis program.

This review of the available programs in the Netherlands strongly suggests that the development of a chopping- and disambiguation procedure for Dutch texts is required to avoid the very cumbersome and unelegant work with KWIC indexes.
With respect to automatic syntactic analysis of Dutch important work is done by M. Boot. As syntactic analysis is critical for disambiguation,

his findings will surely contribute to the solution of problems of computer aided content analysis of Dutch documents. Boot's article in this collection contains further information.

Notes

1) P. Stone e.a. (1966) and for the augmented version D.R. Coxon, A.D. Chalmers (1974)

2) H.P. Iker, N.I. Harway (1969), pp.381

3)E. Kelly, P. Stone (1975), p.3

4) Ibidem pp.8

5)D.R. Coxon, A.D. Chalmers (1974), pp.6

6)H.P. Iker, N.I. Harway (1969), pp.381

7)Critiques of these procedures which mainly concern the validity of the dictionaries, the assumption about frequencies, and the use of factor analysis, can be found in G. Gerbner (1969), pp.523; J. Ritsert (1972), pp.19; etc.

8)A chopping procedure for Dutch has to be developed. However Dutch morphology is much more complex than English. Therefore, solutions will have to be developed, in particular for the handling of syntactical dispersion by which a prefix is split from the verb thus becoming a preposition, etc.

Bibliography

D.R. Coxon, A. D. Chalmers,The General Inquirer, introduction to a computer based system of content analysis, unpublished report, Edinburgh 1974, available at the General Inquirer project, dept. of Sociology, University of Edinburgh

G. Gerbner a.o., The analysis of communication content, New York 1969, J. Wiley & Sons

H.P. Iker, N.I. Harway, A computer systems approach toward the recognition and analysis of content, in G. Gerbner a.o. (1969) pp. 381

J. Ritsert,Inhaltsanalyse und Ideologiekritik, ein Versuch über kritische Sozialforschung, Athenäum Fischer Taschenbuchverlag, Frankfurt am Main, 1972

I.N. Saris - Gallhofer, W.E. Saris, Recente ontwikkelingen op het gebied van de geautomatiseerde inhoudsanalyse, in Acta Politica, no. 754, 1975, pp.485

P. Stone a.o., The General Inquirer: a computer approach to content analysis, M.I.T. Press, Cambridge, 1966