Computer-Assisted Content Analysis

Mergenthaler, Erhard

Veröffentlichungsversion / Published Version
Sammelwerksbeitrag / collection article

COMPUTER-ASSISTED CONTENT ANALYSIS

ERHARD MERGENTHALER

This paper provides an overview of the current state of the art in computer-assisted content analysis (CACA). First, background, history and a model of CACA will be given, the dichotomy of qualitative versus quantitative is addressed, and a new understanding, the "marker view" leading to a more general Text Analysis is introduced. Subsequent chapters provide a definition of terms and cover issues of size of scoring units, and the development of computerized coding to replace well established manual rating systems. The paper concludes with the description of a recently developed computer-assisted text analysis methodology to describe psychotherapeutic processes.

1. Introduction

1.1 History of Computer-Assisted Content Analysis

The method for Computer-Assisted Content Analysis most often is seen as a "mechanized" or "automated" variant of the theme analysis. According to Merten theme analysis is one of the oldest and most wide spread variant of at least 20 conventional content analytic techniques as he more than ten years ago already identified in a synopsis (Merten, 1983:120). As we shall see later in this paper, computer assistance may also be useful, and in fact has been utilized, for many of the other variants as well by now. The terms "mechanized" and "automated" have been put into quotation marks to emphasize rather the wish than the reality of what working with computers can achieve.

Content analysis in general (Früh, 1991; Gerbner et al., 1969; Krippendorf, 1981; Merten, 1983) has been defined as "any research technique for making inferences by systematically and objectively identifying specified characteristics within text" (Stone et al., 1966:5). Theme analysis in special relies on the representational model (Osgood, 1959), roughly saying that manifest texts represent a reliable correlate of the context and therefore it is possible to infer from text to context. Another basic assumption of content analysis is to expect a theme to be the more prominent in a text, the more references can be found to it. This however has been subject of a controversy which has been opened in

1952 by Berelson and Kracauer and which focused on aspects of quantitative vs. qualitative content analysis (Howe, 1988). Although this contention never has been solved finally, it has been seen as being less and less important in the following years, by taking a pragmatic stance. Kracauer, 1972 and later on Howe, 1988 pointed out that the two approaches overlap, with quantitative analyses ending up with qualitative considerations, and qualitative analyses often requiring quantification. Also arguing became less critical due to the convincing results especially computer-aided content analyses brought about more and more often. Dictionaries with many categories were developed and applied in order to assess the manifest content of a text. Well known and probably most often used are the Lasswell Value Dictionary, the Harvard Psycho-Sociological Dictionary (Gerbner et al., 1969; Stone et al., 1966), and the Regressive Imagery Dictionary (Martindale, 1978; 1986; 1990).

The system General Inquirer (Stone et al., 1966) was developed as a prototype for computer-aided content analysis and achieved a position of special importance in the social sciences. Regarding the psychotherapeutic scope of application, Laffal's "total content analysis" (1968) can be mentioned. Although the methodological foundations were distinctly different in the work of Stone et al. and Laffal, the computerized text analysis process followed the same schema for both. The preliminary task is the compilation of a glossary or dictionary, often consisting of a collection of several thousand word forms which are assigned to different categories. The categories themselves constitute a system including either the facets to a special topic or the aspects of a more general complex of topics. The vocabulary of a dictionary can be derived either *inductively* from a text or *deductively* from more general constructs whose consequences can be detected in the choice of categories. The computer's task is to examine a text word for word and to compare it to the dictionary. If a word form is found, the number of entries counted for the corresponding category is increased by one. The resulting frequency distribution can also be relativized according to the text for the purposes of comparison. Depending on the system, this fundamental algorithm can be modified into a more or less elaborate form by the introduction of additional rules. Stone et al., for example, attempt in this manner to resolve the ambiguity of many word forms by referring to the context. This level of development characterizes computer-aided content analysis even today, including the newer systems commonly used today (TEXTPACK, Intext, TAS). It can be considered an independent method that, however, barely has gone beyond the scope of its application in the empirical social sciences. As late as 1993 Ray Siemens published a paper, proposing practical content analysis techniques for text-retrieval in large, un-tagged text-bases.

The "golden age" of computer-assisted content analysis is marked by the years from 1960 to 1970. Within numerous contributions - merely all of them within the Anglo-American literature working mainly on mass communication research and literary text analysis - methodological implications have been discussed and basic applications have been shown (e.g. Gerbner et al., 1969; Laffal, 1968; Stone et al., 1966). The years from 1970 to 1980, although the method also was discovered by psychotherapy researchers (Dahl, 1972; Spence, 1970), and also became more wide spread in Europe, are characterized by a slowing down of published reports on specific applications in the primary fields of applications. But never this technique was totally out of use. Nowadays we may see computer-aided content analysis as a standard methodology within the empirical social sciences including psychotherapy (Rosenberg et al., 1990), but also as a methodology that shows methodological stagnancy since years. Due to a variety of overviews which appeared in recent years however, there is an increasing understanding of inherent problems like reliability, validity (Weber, 1983), statistical issues (Hogenraad & Bestgen, 1989; Hogenraad et al., 1995), linguistic aspects (Frühlau, 1981; Jeanneau, 1991) and practice oriented guidelines (Weber, 1985; Züll et al., 1991).

A totally different but also computer-assisted approach using statistical methods to find categories, Harway & Iker (1964) made use of. This technique did not find however further attention.

1.2 A Model for Computer-Assisted Content Analysis

The following model is given in order to make the process of text analysis transparent. It starts from a bipartite view of a real and a formal system. A natural language is postulated within the real system and a formal language within the formal system.

Furthermore, the real system is divided into an object-linguistic and a meta-linguistic component. Any text that will be analyzed is now interpreted as an object-linguistic realization within the real system. The guiding theory for the text analytic process is handled as a meta-linguistic component. The formal system comprises a category system without any further differentiation. The procedural description of the text analysis now can be given in three steps:

1) Translation of a text from the real system into a formal category system.

2) Interpretation of the formal category system within a theory.

3) Evaluation and verification of the findings with the text being analyzed.

This model is appropriate in the description of scaling techniques for verbal material as for example, the anxiety scales of Gottschalk and Gleser (1969) or Bucci's Referential Activity (Bucci & Kabasakalian-McKay, 1992).

By means of computer-assisted text analysis the crucial work of translation as a first step is performed by computational rules implemented as part of the software used. The second step involves a coding procedure. Within the model this results in a further differentiation of the formal system into object-linguistic and meta-linguistic components (see figure 1). Thus there will be a correspondence between text and extracted information (e.g. vocabulary), and theory and category system respectively. Procedural description now comprises four steps:

1) Reduction of a text to selected information.

2) Translation of the information to a category system.

3) Interpretation of a category system within a theory.

4) Evaluation and verification of the findings with the text being analyzed.

While the first two steps now are performed mostly automatically using text analysis software, the third and fourth steps are based on human skills, making use of the complex contextual environment and universe of meaning.

The use of quantitative methods in general implies, according to the above model, a very rigid reduction of the variety of information originally present in a text. In contrast to certain hermeneutic approaches, in which ideas and additional interpretative information may even be added, computer-aided text analyses build to the more general phenomena. According to a generally accepted understanding of science, the goal of research must be to find agreement among scientific communities which ultimately can lead us to more generally accepted knowledge. From the hermeneutic point of view, if the object and the applied methodology are too complex to be understood and accepted by others, the researcher tends to look for refinement using smaller and more readily judgeable objects and methods. In case of computer-aided text analysis the research strategy is just the opposite. If agreement can't be found, stepwise generalization will be sought until consensus is found.

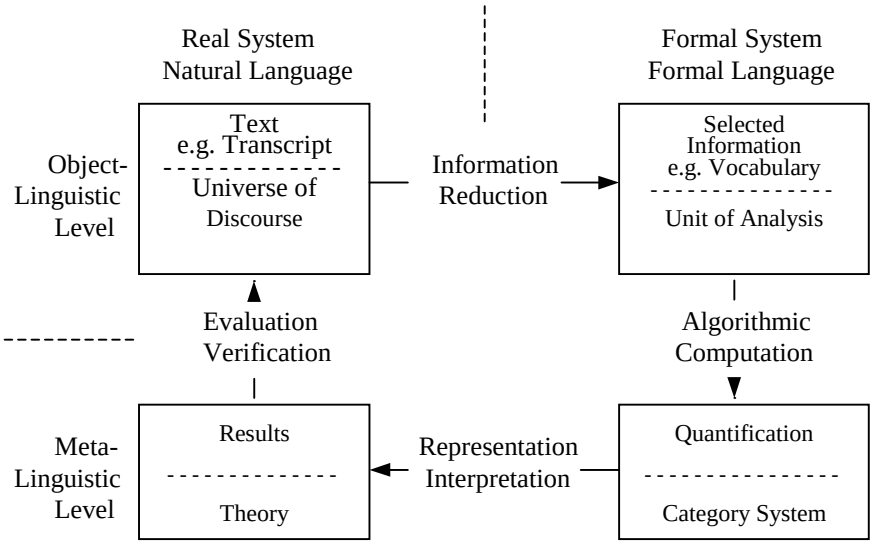

Figure 1: Model of computer-assisted content analysis

1.3 Category Systems - Different Views

As it may be obvious the decisive tool in computer-aided content analysis resides with the dictionary as it connects vocabulary with theory. A dictionary is defined as both: *wordlist* and *category system* and thus corresponds to the formal component in the above mentioned model.

The Selective View

Stone et al. (1966) differentiated specific and general dictionaries. A *specific* dictionary serves as an instrument for the investigation of a narrow and well defined problem. For example, we refer to the Anxiety Theme Dictionary developed at Ulm University (Grünzig, 1980; Grünzig, 1983; Speidel, 1979). It comprises four categories called Shame, Mutilation, Guilt, and Separation. A *general* dictionary serves as a tool in the investigation of various not necessarily predefined problems. A well known example may be the Harvard Third Psychosociological Dictionary with its 53 categories. At a closer look, however, generality in this example reveals as the composition of several specific dictionaries into the general frame of objects, processes, and qualifiers.

The Total Content Analysis View

A quite different view of generality is proposed by Laffal (1968). He promoted a Conceptual Dictionary for use within a "total content analysis" of language. The category system should be highly dependent from the cognitive capabilities and experiences of human beings. In a comprehensive rationale Laffal makes use of specific reading like Piaget, Vygotzkij or Hallig and Wartburg. His conceptual dictionary comprises 114 categories.

Laffal´s Total Content Analysis implies the coding of almost every textword except those with extreme high frequencies in normal language, as we find with function words. This contrasts with Stone et al., who list the 5000 most frequent words to use for dictionary construction. This results in different types of content analysis. The former picks up a large variety of highly content-dependent nouns, adjectives, and verbs. The latter deals with everyday vocabulary. On the other hand this results in a text coverage of 10 percent for Stone et al. and of 90 percent according to Laffal.

The Marker View of Content Analysis

While the dictionaries presented so far in this paper were understood as directly revealing the contents of some text, emphasis now is put on a use of dictionaries that indirectly reveals the contents of the analyzed text. In contrast to the former view this does not intend to quantify the manifest content of a text but rather to identify more general thematic aspects. An entry from the dictionary, which matches a word in the analyzed text is seen as a "marker" indicating the presence of a thematic construct but not really measuring it. Given this approach the controversy about qualitative and quantitative aspects is made obsolete by identifying phenomena rather than reporting categorical data. Computer-aided text analysis as it is proposed along with the marker view will provide us with the following information: Are the phenomena we are looking for present in the text, and if so, where are they located? However text analysis will not tell us the specific variant or the intensity of a phenomenon in the text.

2. Definition of terms

The following list of useful definitions is included as an aid to the reader. Such a definition of terms has proven itself useful in the past since the everyday meaning of some of the terms diverges from the special meaning in the field of computers.

Word Form

A word form is every word written by an author or said by a speaker of a natural language. In the written presentation of speech, word form refers to a sequence of letters bordered by spaces or special symbols. Examples of word forms are the words: I, gone, houses, hm.

Comment: In general linguistics this definition corresponds to a series of graphemes. There a word form can consist of several, not necessarily sequential, series of graphemes; for example, "will have eaten" is a word form consisting of three groups of graphemes. In a sentence this word form may be interrupted by other graphemic groups: "Tomorrow I will only have eaten lunch or had a snack." The recognition of such word forms with the aid of algorithmic procedures requires extensive syntactic and semantic analyses of the context. In many cases, the end of the sentence does not sufficiently limit the context, so that the entire text may possibly have to be referred to.

Basic Form

Basic forms are all uninflected word forms. For verbs this is the infinitive, for nouns the nominative singular, and for adjectives the positive. Thus for one basic form there may be several word forms. Examples of basic forms are the words: I, go, house, hm.

Comment: Corresponding to the definition of word form used here, the concept of basic form also refers to a series of graphemes: "will have eaten" thus refers to the three basic forms "will", "have", and "eat". In general linguistics this would simply be "eat".

Complete Form

A complete form is an inflected word form, and for every complete form there is a basic form. Examples of complete forms are: me, went, houses. Thus the above defined term *word form* does not differentiate between *basic* respectively *complete* forms.

Word List

A word list is a finite quantity of different word forms.

Lemma

A lemma is a quantity of word forms grouped together because of their agreement with regard to given qualities. Examples of lemmata are the word entries in a standard dictionary like Webster's.

Comment: General linguistics distinguishes between syntagmatic, paradigmatic, and structural qualities. A second condition is usually that a lemma only includes word forms of one part of speech and with one root.

Lemma List

A lemma list is a structured quantity of lemmata. Every single lemma is determined by a congruence of the word's part of speech and the meaning of all the word forms belonging to it. The corresponding basic form is used as the lemma name. Examples for the entries in a lemma dictionary are the lemmata:

residence = (residence, residences)

reside = (reside, resided, residing, resides)

Part of Speech

Part of speech is the role a word form fulfills in speech or in sentence structure (see Erben, 1968:38ff.). The following parts of speech are distinguished:

- Verbs
- Nouns
- Adjectives
- Pronouns
- Prepositions and Conjunctions
- Adverbs and Predicate Adjectives
- Interjections

Verbs amount to about a fourth of the entire vocabulary and are the main means of statements (rheme) describing action or a state of being. Nouns constitute more than two-thirds of the entire vocabulary and serve to name the significant aspects (theme) surrounding an action or determining a state of being. Adjectives constitute about a sixth of the entire vocabulary and serve to characterize a given act or state of being and the significant aspects involved in it. The pronouns, prepositions, conjunctions, and adverbs together amount to about a tenth of the entire vocabulary; their function is to supplement the three main word kinds by enabling references, relationships, connections, and modal and emotional expressions to be made.

Comment: This functionally and syntactically oriented definition of part of speech takes especially the pragmatic goals of the desired text analysis into account.

Word Form Index

A word form index is a structured list of all the word forms appearing in a text. The frequency of a word's appearance is also recorded for each word form.

Basic Form Index

A basic form index is a list of all the word forms appearing in a text which have been traced back to their basic forms. The frequency of a word's appearance is recorded for each basic form.

Comment: Word and basic form indexes constitute the basis of frequency dictionaries. A comprehensive description is given by Alekseev (1984).

Category

A category refers to the names of open quantities of word forms grouped under the same substantive point of view. For example, all word forms which refer to sensual perceptions could be grouped under the category "sense".

System of Categories

A system of categories is a quantity of categories which is self-contained according to substantive points of view. For instance, the system of categories ANXIETY THEMES includes the categories Shame, Mutilation, Guilt, and Separation.

Vocabulary

Vocabulary refers to an index of complete forms or of basic forms if the distinction between complete form and basic form is not relevant.

Dictionary

A dictionary is a finite quantity of ordered pairs. Every pair consists of a word form and a category. For example, given the word list:

W = (cut_off, cutting_off, soon, knife, judgement, judgements)

and the category system

C = (Shame, Mutilation, Guilt, Separation).

A dictionary which corresponds to the definition given here might look as follows:

D = (cut_off, Mutilation; cutting_off, Mutilation; cut_off, Separation; cutting_off Separation; knife, Mutilation; knife, Guilt; judgement, Guilt; judgements, Guilt)

Such a dictionary can also be understood as a relation between word list W and the system of categories C, and thus as a subset of the Cartesian product W x C. To obtain the desired sub quantity in the form of a dictionary, secondary conditions can be agreed upon. The most common ones are:

1. For every pair of elements in subset D, the meaning of the given word forms should agree with the definition of the category associated with them.

This excludes the "meaningless" pairs of elements from the complete Cartesian product. Since this condition was applied to the above-mentioned example, the pair of elements (soon, Guilt) is not given there as an item in the dictionary D.

2. Each word form in the word list W may appear in only one pair of elements in the dictionary D.

This agreement prevents multiple classifications. If the second condition is applied to the above-mentioned example, a decision must be made as to the category under which the word forms "cut_off", "cutting_off", and "knife" will be included in the dictionary. Consequently the following dictionary might result:

D = (cut off, Mutilation; cutting off, Mutilation; knife, Mutilation; judgement, Guilt; judgements, Guilt)

3. Excluding inflected word forms makes dictionaries smaller and easier to use.

Applying the third condition to the example produces the following dictionary:

D = (cut off, Mutilation; knife, Mutilation; judgement, Guilt)

Standard Dictionary

Standard dictionaries are defined as dictionaries satisfying all three secondary conditions.

Text

A text is a structured quantity of word forms, punctuation marks, and commentaries. Symbols identifying speakers, chapters, or other structural information do not belong to the text itself, but label a text. A text can be organized hierarchically. Typical levels that are distinguished are:

- Word form
- Utterance, Paragraph, Statement
- Session (hour), Chapter
- Sequence of sessions (treatment), Book

Standard text

A standard text is a text whose word forms have been traced back to basic forms. A standard text thus does not contain any inflected forms, but its structure (and therefore the sequence of the word forms) is retained.

Type

All the different word forms appearing in a text or a word list are called types.

Token

All the word forms appearing in a text or word list are called tokens. The number of tokens always corresponds to the text size.

Corpus

A corpus is a quantity of text which is grouped together under one general point of view. For example, the Ulm Textbank comprises a corpus which contains texts form the psychotherapeutic situation. Subcorpora can also be defined; for instance, the collection of first interviews selected on the basis of the patient's sex and age constitutes a limited subcorpus.

Processing

The algorithmic processing of texts is possible according to two points of view. First, a text can be viewed as a set of word forms and processed according to a set-oriented procedure. The other view follows the sequential structure of word forms in the text; such procedures are called structure oriented.

Set-Oriented Processing

Preparation of a Word Form Index

Word forms, together with frequency of occurrence, which appear in a text are determined, sorted either alphabetically or according to frequency of occurrence, and made available as a file or a printout.

Preparation of a Basic Form Index

A basic form index can be prepared from either a text or a previously prepared word form index. All the complete forms which occur are traced back to their basic forms (standard text), sorted again either alphabetically or according to frequency of occurrence, and made available as a file or a printout.

Difference Between Vocabularies A and B

Vocabulary X is determined. It consists of all pairs of elements contained in vocabulary A but not in vocabulary B.

For example, given the vocabularies

A = (I, 3; you, 7; he, 5; she, 8) and

B = (she, 2; we, 1)

Vocabulary C then contains C = (I, 3; you, 7; he, 5).

Furthermore, it is also possible to determine the *limited difference* between the vocabularies. In other words, vocabulary X may include pairs of elements which have the same word forms and which appear in both vocabularies if the ratio between the frequency of a word's occurrence in vocabulary B to that in vocabulary A does not exceed a specified value. Thus in the example, the pair (she, 8) belongs to vocabulary X given a limiting value of R = 0.25.

Intersection of Vocabularies A and B

Vocabulary X is determined. It consists of all pairs of elements present in both vocabularies A and B (possibly with different frequencies of occurrence). The frequency for a pair in vocabulary X is the sum of the frequencies in A and B.

Characteristic Vocabulary of Two (or More) Texts A and B

Vocabulary X_A is determined. It consists of all pairs of elements being significantly more frequent present in text A than in text B. The level of significance can be determined (usually probability of error is set p = 5%).

Application of Dictionary D to Vocabulary A

Applying dictionary D to vocabulary A produces a distribution of the categories in vocabulary A and an internal differentiation of each category. Using the concepts "relation" and "selection" which were taken from relation algebra following Codd (1970) makes it possible to present the processing forms described here in a formal

mathematical manner (not shown in detail here for the sake of clarity). For example, with dictionary D = (cut_off, Mutilation; knife, Mutilation; judgement, Guilt) and vocabulary A = (cut_off, 4; knife, 4; judgement, 2) we have the relation X = (cut_off, Mutilation, 4; knife, Mutilation, 4; judgement, Guilt, 2). This produces the selections S_1 = (Mutilation, 8) and S_2 = (Guilt, 2). This example is presented in Table 1.

Word Form	Category	Frequency	Type
cut_off	MUTILATION	4	Relation X_1
knife	MUTILATION	4	Relation X_2
	MUTILATION	8	Selection S_1
judgement	GUILT	2	Relation X_3
	GUILT	2	Selection S_2

Table 1: Example of the application of a dictionary to a vocabulary

Application of Two Dictionaries One to Each Other

Applying a dictionary D_1 to a dictionary D_2 determines a relation X that explains which categories in dictionary D_1 measure the categories in dictionary D_2. An example is given in table 2.

Structure-Oriented Processing

Application of a Dictionary to a Text

Every word form in a text is searched in a dictionary and replaced by the appropriate category. Word forms not contained in the dictionary are replaced by the category "undefined". The product of this form of processing is a sequence of categories corresponding to the sequential structure of the text.

Processing a Sequence of Categories

With the aid of conditions and instructions it is possible to manipulate a sequence of categories. The following conditions are possible:

- The occurrence ("exists") of a category within the sequence studied
- The position of a category within the sequence studied
- The relation to the categories preceding and following a category within the sequence studied
- The conjunction of two categories within the sequence studied

D_1	Word Form	Category	D_2	Word From	Category
	accuse	ATTACK LEGAL SIGN-REJECT		accuse	GUILT
	doctor	AUTH-THEME HIGHER STATUS MEDICAL		doctor	MUTILATION
	mother	FAMILY FEMALE-ROLE HIGHER STATUS		mother	SEPARATION
	bankruptcy	AVOID ECONOMIC		bankruptcy	SHAME
	rival	ASCEND-THEME ATTEMPT SIGN-REJECT		rival	MUTILATION
	worry	DISTRESS SIGN-WEAK		worry	GUILT
	home	FEMALE THEME SOCIAL PLACE		home	SEPARATION
	naked	DANGER-THEME SENSORY-REFERENCE SEX-THEME		naked	SHAME

X	D_2	D_1
	Category	Category
	SHAME	AVOID
		ECONOMIC
		DANGER THEME
		SENSORY REFERENCE
		SEX THEME
	MUTILATION	ASCEND THEME
		ATTEMPT
		AUTH THEME
		HIGHER STATUS
		JOB ROLE
		MEDICAL
		SIGN REJECT
	GUILT	ATTACK
		DISTRESS
		LEGAL
		SIGN REJECT
		SIGN WEAK
	SEPARATION	FAMILY
		FEMALE ROLE
		FEMALE THEME
		HIGHER STATUS
		SOCIAL PLACE

Table 2: Example for the evaluation of two dictionaries.

Where such conditions apply, one of the following instructions can be carried out and stored in a file or become printed:

- Addition of a category
- Deletion of a category
- Substitution of a category
- Selecting of the text preceding and following a category

Interaction Sequences

Starting from a sequence of categories, it is possible to formalize the sequences of categories for a change across analysis units and to analyze them with specially prepared models.

Summary of the Forms of Processing

It is possible to divide the methods for analyzing texts into two groups by distinguishing between set- and structure-oriented forms. Another criterion for classifying the methods is given by the possibility to analyze texts resulting from a conversational situation according to a monadic or dyadic approach.

In the monadic form, only the contribution of an individual speaker or author is used in the analysis. In the dyadic form, in contrast, especially those speech phenomena are taken into consideration that result from the interaction of all participating speakers. Thus four basic groups of methods can be distinguished in computer-assisted text analysis:

	MONADIC	DYADIC
SET ORIENTED	Group 1	Group 2
STRUCTURE ORIENTED	Group 3	Group 4

The choice of one of these groups of methods depends on the goal of the research. In the following, several examples of applications are listed together with the group of methods which is especially appropriate.

Group 1

Alterations in a subject's speech across time

Comparison of different speech situations

Comparison of different groups of speakers

Group 2

Interaction sequences in the studied dyad

Speaker typologies

Group 3

Structures of a subject's associations

Group 4

A subject's communicative strategies

Generally, methods from different groups are combined for more extensive kinds of questions in order to get results which are more reliable.

3. Some Technical Aspects

3.1 Size of Scoring Unit

Minimal Text Size. If text data are to be used in a scientific study to determine measurements considering both frequency and distribution a text needs to be divided into several segments. Making separate measurements on each section, the question is quickly posed as to the determination of the appropriate sample size. This aspect concerns the length of a section of text used in the analysis. This can be of significance, for example, in the decision as to whether single paragraphs/utterances or entire chapters/conversations should be used. There have not been any sound scientific studies on these questions. As far as the text size has been taken into consideration in published studies at all, the authors rely primarily on the observations that they were able to make, while analyzing speech material. For conventional content analysis Gottschalk and Gleser (1969) and Gottschalk et al. (1969) for example, determine that their anxiety scales could be utilized for a text of at least 70 (English) words. Schöfer (1977) gives 100 words as the minimum value for the German version of these scales. This minimum is founded methodologically in the reliability with which numerous judges were able to classify test sentences. Ruoff (1973) describes texts of 200 word forms as the minimum size which is also applicable in practice. He bases this on his view that phenomena necessary to speech are normally distributed; this normal distribution begins to occur in the corpus of spoken speech that he studied in southern Germany at this text length. For computer-assisted text analysis there are no well-founded indications of minimum size; usually values between 500 and 1000 word forms are suggested without further discussion.

As a significant aid to practical work, especially to computer-assisted content analysis the following formula to estimate the necessary sample size N for the methods mentioned above is given by Mergenthaler (1985).

$$N > \ln(\alpha/2) / \ln(1-p)$$

where α is the level of significance the analyses are based on (usually $\alpha = 0.05$) and p is the expected frequency as determined from the basic vocabulary for the category with the least frequency within a given dictionary (ln = logarithm base 2).

The considerations leading to the formula start from the idea that the process of speech or text production in man can be understood as a stochastic process (Bennett, 1977). In other words, laws of probability describe the origin of a text and determine the sequence of the individual word forms in it. Although these laws are at first unknown, they can be discovered empirically. The following assumption is therefore made: The sequence of all the word forms in a collection of texts, called text corpus for short, represents the laws of probability.

It is possible on this basis to determine the basic vocabulary of a text corpus. Statistically, the basic vocabulary can be understood as the parent population from which a concrete text is taken as a sample. The frequency of occurrence of individual word forms contained in the basic vocabulary can be viewed as their general probability of occurrence; it thus constitutes an empirically determined measurement of probability. The idea of a basic vocabulary is also used by Henken (1976) in a computer-aided content analysis of documents from people who attempted suicide. He comments with regard to sample size: "Each group contained a minimum of 1000 words to insure reliability" (p. 37). The values for different categories, determined using the Harvard Psycho-Sociological Dictionary, he compared with a baseline derived from a one-million-word corpus of American prose (Kucera & Francis, 1967). The values he got for the proportions of the categories range between 0.06% and 8.93%. To be able to compare the results for individual texts, the significant deviations from the baseline in both directions were determined at various levels of probability of error. Not taken into consideration, however, was whether a significant deviation from the given probability of error is at all possible for a sample size of 1000 words and a, for example, 0.06% general probability. The formula above, demonstrates that the minimal sample size is 6146 words at a 5% probability of error for the category "medical", which has a general probability of 0.06%. Thus the 1000 words chosen by Henken are not sufficient to permit the findings to be interpreted. Table 3 gives the minimal text size for some sample probabilities.

p	N
0.01	36.886
0.05	7.375
0.10	3.687
0.50	735
1.00	367
5.00	71
10.00	35

Table 3: Minimal text size N for sample probabilities p of the expected frequency of occurrence of a category with a given level of error of alpha = 5%.

3.2 From manual rating to computerized coding

The following section deals with a problem that is very common: How to develop a computer-assisted procedure for scoring textual material in order to replace a well established manual rating system? There are only few systematic studies dealing with this topic and comparing human ratings with computer attempts for the same scales or categories. Some of them focus around the computerization of the Gottschalk & Gleser Anxiety Scales (Gottschalk & Gleser, 1969; Gottschalk et al., 1969). Gottschalk and Bechtel presented an attempt to measure these scales by modelling the rater's steps of segmenting and coding as close as possible on a computer (Gottschalk & Bechtel, 1982). The results however have not been very satisfying. Partly this was due to the limited computational linguistic knowledge and tools available at that time. Meanwhile Gottschalk and Bechtel (1995) have a system that makes use of a linguistic parser and a knowledge based system. The results are reported to be comparable to those of human raters. The efforts to develop this system took years and success only was possible by collaborating with computer scientists and computational linguists.

An attempt more in line with the technique of computer-assisted content analysis was made by Speidel (1979). In a deductive approach she developed a specific dictionary having one category for each of the Gottschalk & Gleser Anxiety Scales. Grünzig and Mergenthaler (1986) then compared results from judges with the computer based measurements and had to realize that although many convergent findings, there were also instances with significant differences. They conclude that the Anxiety Theme Dictionary measures something different than the original rating scales. This does not mean however that the computer approach is useless. In contrary, it has proven to be a valuable tool in psychotherapy process research.

More recently Mergenthaler and Bucci (in prep.) presented a new technique in modeling the Referential Activity rating scales (Bucci & Kabasakalian-McKay, 1992). Referential Activity (RA) is the amount of active links between the verbal and non-verbal system of a person. This can be observed in textual data. The rating is done on four eleven point scales: Concreteness, Imagery, Specificity, and Clarity. An overall measure of RA is given by the mean over all four scales. RA usually peaks when a narrative is encountered in a text. Mergenthaler and Bucci now made use of the expertise human raters have. They took rated material and selected extreme samples from both ends of the scale (0-3 and 7-10). Thus they got two corpora, one with prototypic low RA text and one with high RA text. In a next step the *characteristic vocabularies* (see definition on page 14) were computed for these corpora. The one vocabulary resulted with words which are considered to be typical for low RA, the other one with words being characteristic for high RA. After minor editing the lists were ready as a dictionary for Computer Referential Activity (CRA) comprising two categories, one for high and one for low RA. The correlation coefficients for CRA with human raters was found to be within .42 and .65 which is considered to be very good. In fact, treating the computer as a rater, the inter-rater reliability is within the range of human raters for CRA.

4. Example of a Computer-Assisted Text Analysis Methodology

In this chapter, a study will be presented that has been published by Mergenthaler (in press) and which had the goal to develop a computer-aided system that is able to identify key moments in transcripts from psychotherapeutic sessions. The term «key moment» refers to one or more sessions of a treatment or to segments within a session which are seen as clinically important. These include moments which may be seen as a turning point or break through, moments which mirror points of insight as they occur in the course of the psychotherapeutic process and which are needed in order for some change in the patient's demeanor to take place.

4.1 Background

The method is based on two variables. The one is emotion, a phenomenon that is seen as a central aspect for many or all psychotherapies. The concept «emotion» is understood as «emotional tone of a text», a term as it is used also in literary and linguistic research (Anderson & McMaster, 1982; Anderson & McMaster, 1986; Anderson & McMaster, 1990). Thus utterances or words will be observed that are suitable to verbally express

emotion, which however may not coincide with physiological correlates like sweating, flushing, or palpitation.

As a second variable «Abstraction», a construct leading to the development of understanding and perception (Piaget, 1977) and thus patients' main cognitive activity, was chosen. Abstraction has clearly observable linguistic effects. Besides a rich resource of abstract nouns natural language provides the unlimited possibility to build abstract terms out of concrete concepts by performing a morphological transformation on single word forms, like from "tender" to "tender-ness".

Both, Emotion Tone and Abstraction are assumed to vary in intensity during the therapeutic process. Furthermore it is expected that the possible combinations of Emotion Tone and Abstraction as expressed in language have clinical significance. This leads to the following general hypothesis: For a «good hour» (key session) to emerge, the temporal coincidence of Abstraction and Emotion Tone is a necessary condition. The same is true for a «good moment» (key moment) within a session. For the empirical assessment four states based on the notion of Emotion Tone and Abstraction will be defined and introduced as «Emotion-Abstraction Patterns». The four patterns are defined, labeled, and interpreted as follows:

Pattern A - Relaxing: Little Emotion Tone and little Abstraction. Patients talk about material that is not manifestly connected to their central symptoms or issues. Their stance of speaking is rather describing than reflecting. Also it is a state where patients return to as often as they feel the need to regenerate both, physis and psyche to prepare themselves for the next step of their «talking cure».

Pattern B - Reflecting: Little Emotion Tone and much Abstraction.

Patients present topics with a high amount of abstraction and without intervening emotions. This may be an expression of defense known as intellectualizing.

Pattern C - Experiencing: Much Emotion and little Abstraction.

Patients find themselves in a state of emotional experiencing. Patients may be raising conflictual themes and experiencing them emotionally.

Pattern D - Connecting: Much Emotion Tone and much Abstraction.

Patients have found emotional access to conflictual themes and they can reflect upon them. This state marks a clinically important moment, this is the instant that was introduced as key moment earlier.

The following model is derived from a specific temporal sequence of the four Emotion-Abstraction Patterns. This will be introduced as «Therapeutic Cycle» consisting of five phases. It is based on the assumption, that in the course of a psychotherapy or within a psychotherapy session Emotion-Abstraction Patterns do not occur by chance. Rather a periodic process for the underlying variables Emotion Tone and Abstraction is assumed. To explain this not only psychic, but also biological factors may contribute (e.g. endorphins). A good example of how humans are used to, how they even need a behaviour as such, Johnson has given with the following anecdote:

«A harpist who lingers too long on one string offends our ear; just so, the speaker who remains too long on the same general level of abstraction offends our evaluative processes–no matter what his subject may be.

The story is told of the man who played the bass viol. But he didn't play it the way other people play a bass viol. His bass viol had only one string, and he kept his finger always in the same place while he bowed that one string. In this way he played long, long, day after day–until his wife became exasperated, gentle soul though she was. "John," she said, "why don't you play the bass viol the way other people do? Haven't you noticed that they have many strings on their bass viols, and they move their fingers up and down all the time when they play?"

"Sure they do," said John, as he went on bowing. "They're looking for the place. I've found it."» (Johnson, 1946:278)

Phase I: Starting point is pattern A (Relaxing), moments where patients do not show much emotion nor abstraction. They find themselves in a "relaxed" state, in a transitional state from one theme to another, or they are associating freely.

Phase II: After a while emotion increases and pattern C (Experiencing) will show up. This shift can be initiated by having reported a narrative (dream, early memory, episode) or by reporting on the symptoms they are suffering from. Patients at this time are in a state of emotional experience.

Phase III: Ideally then the amount of reflecting will increase, either by patients' own impetus or guided by the therapist. Patients will reflect their recent emotional experience and thus reach at emotional insight. They are in a state of connecting Emotion Tone and Abstraction showing up as pattern D (Connecting).

Phase IV: As a consequence of the insight processes the emotional tension will decrease. Patients can reflect upon their new experience without being bound to emotional constraints. Pattern B (Reflecting) will show up.

Phase V: Finally reflection will fade out as well. The cycle ends with the state of Relaxing (pattern A) which shortly after can lead to the emergence of a new cycle.

The Therapeutic Cycles Model allows for both, a *macro-analytic* view over the course of a treatment, and for a *micro-analytic* view describing the flow within a session.

Macroanalysis: If the Emotion-Abstraction Patterns are computed for complete therapy sessions a therapy can be characterized by the given sequence of these patterns. Turning points are given by the session immediately before a shift into a new pattern. Key sessions will show up with the pattern D (Connecting).

Microanalysis: The Therapeutic Cycle Model describes the very moments of genesis, effect, and end of therapeutic progress. It is not expected to find that the Therapeutic Cycle occurs frequently or repeatedly within a session nor to find one in every session.

4.2 The Development of the Dictionaries

The Emotion Tone dictionary was compiled from various word lists taken from literature and from a body of approximately one million words of running text, taken from English texts stored in the Ulm Textbank, which were examined for emotionally tinged words. The developed word list was revised in such a way that words with concrete aspects of sensory reference as for example "heart" or "warm" were deleted from the dictionary, and words which could not be classified into at least one of the following dimensions: PLEASURE-DISPLEASURE; APPROVAL-DISAPPROVAL; ATTACHMENT-DISATTACHMENT; SURPRISE.

The final dictionary comprises only a single category with the common term "Emotion Tone" (ET), and does not use any further divisions. In its current form, the ET dictionary consists of 2305 items, including inflected forms. In a sample of 80 sessions from 20 patients ET covers an average of 5.4% of the text, with a standard deviation of .62%.

The Abstraction dictionary was obtained primarily through a suffix analysis of all words in English texts which were available in the Ulm Textbank. This technique goes back to the examination of Gillie (1957), who showed that the use of specific endings (e.g. -ness, -ity), which is typical for abstract word forms, correlates significantly with the classification of texts by observers regarding the construct of abstraction.

4.3 Textual Data Used

The method of measuring Emotion-Abstraction Patterns was tested using two different types of clinical text corpora. The material has been chosen in such a way that the validity of the method can be demonstrated as well. This became possible because independ-

ent clinical evaluations and results from psychological tests are available for both corpora, which we can interpret in relation to the hypotheses that we want to support. The first corpus is a sample of 80 sessions taken from the Penn Psychotherapy Study (Luborsky et al., 1988). The sample consist of four sessions from each of 20 treatments. Ten of these patients had a good outcome, 10 did not improve. The second corpus is a single case covering all 28 sessions of a psychodynamically oriented short term psychotherapy provided by the Project on Conscious and Unconscious Mental Processes (Horowitz et al., 1993). We know from independent clinical studies of this treatment that it had a clear "key session" (no. 12) and within this session also two "key moments".

4.4 Segmenting of transcripts

In order to describe the flow of a variable within a session a segmentation into scoring units is needed. The measurement then can be done for each segment independently and the course of the respective variable can be observed, or further analyzed by use of the sequence of the measured values. For statistical reasons there should not be less than seven to ten scoring units. On the other hand the upper bound is also limited otherwise a single segment would become too small in terms of number of words included (see chapter 3.1 above). This estimate has to be based on the variable with the least text coverage. Abstraction with about 4% thus needs a minimum of 129 words (see also table in Mergenthaler 1985:173). The often used "idea units" or "thought units" (Butterworth, 1980) are not suitable, because they normally just comprise little more than one sentence. Therefore transcripts were segmented into word blocks of 150 words each. Obviously this kind of segmentation will not take care of a thematic flow within the session. To compensate for this the data were smoothed using a weighted mean (1-2-1) spanning over three word blocks.

4.5 Results

The Emotion Tone and Abstraction dictionaries together cover a total of nine to ten percent of the text. The improved patients have significantly higher levels for the variable Abstraction ($p<.05$). For Emotion Tone the difference is not significant, however successful patients have a higher level of emotion here too. As for the Emotion-Abstraction Patterns improved patients had less Reflecting and more Connecting ($p<.10$). With regard to Experiencing and Reflecting the samples did not show a difference. Comparing early with late sessions an increase of Connecting for the improved patients statistically was supported (t-test for paired samples, $p<.05$).

The Single Case analysis clearly revealed the key session which was also identified by the therapist and by a research team. Within this session a Therapeutic Cycle was found. It occurs right in the moment were the most prominent change of the patient took place.

4.6 Conclusion

With the method presented here the clinical concept of emotion is brought together with the linguistic phenomenon of abstraction and shown as being productive for the therapeutic process. It allows to operationalize and to measure the important concept of emotional insight in a transparent way. The Therapeutic Cycle describes the psychodynamic of a psychotherapy.

Figure 2: Emotion Abstraction Patterns and the Therapeutic Cycle in Case 40, Session 340

Figure 2 gives an example for a Therapeutic Cycle which is preceded by two narratives (dream reports) as measured by Computer Referential Activity (CRA). Both times the patient got interrupted by a phone call which was answered by the therapist. As a consequence the patient's resistance increases and only due to an intervention by the therapist Connecting is achieved.

Literature

Alekseev, P. M. (1984): *Statistische Lexikographie. Zur Typologie, Erstellung und Anwendung von Frequenzwörterbüchern***.** Bochum: Studienverlag Dr. N. Brockmeyer.

Anderson, C. W. & McMaster, G. E. (1982): Objective analysis of emotional tone in stories and poems. *Journal of the Association for Literary and Linguistic Computing*, 3:45-51.

Anderson, C. W. & McMaster, G. E. (1986): Modeling emotional tone in stories using tension levels and categorical states. *Computer and the Humanities*, 20:3-9.

Anderson, C. W. & McMaster, G. E. (1990): The emotional tone of foreground lines of poetry in relation to background lines. *Journal of the Association for Literary and Linguistic Computing*, 5:226-228.

Bennett, W. R. (1977): How artificial is intelligence? *American Scientist*, 65:694-702.

Berelson, B. (1952): *Content analysis in communication research*. Glencoe: Free Press.

Bucci, W. & Kabasakalian-McKay, R. (1992): *Scoring referential activity. Instructions for use with transcripts of spoken narrative texts*. Ulm: Ulmer Textbank.

Butterworth, B. (1980): Evidence from pauses in speech. In: Butterworth, B. (Ed.): *Language production* (pp. 155-176). London: Academic Press.

Codd, E. F. (1970): A relational model of data for large shared data banks. *Communications of the American Computing Machinery*, 13 (6):377-387.

Dahl, H. (1972): A quantitative study of a psychoanalysis. In Holt, R. R. & Peterfreund, E. (Eds.): *Psychoanalysis and contemporary science* (pp. 237-257). New York: Macmillan.

Erben, J. (1968): *Deutsche Grammatik - ein Leitfaden*. Frankfurt a. M.: Fischer Taschenbuch Verlag.

Früh, W. (1991): *Inhaltsanalyse, Theorie und Praxis*. München: Ölschläger.

Frühlau, I. (1981): Inhaltsanalyse versus Linguistik. *Analyse und Kritik*, 3:23-41.

Gerbner, G., Holsti, O. R., Krippendorf, K., Paisley, W. J. & Stone, P. J. (1969): *The analysis of communication content. Developments in scientific theories and computer techniques*. New York: John-Wiley.

Gillie, P. A. (1957): A simplified formula for measuring abstraction in writing. *Journal of Applied Psychology*, 41: 214-217.

Gottschalk, L. A. & Bechtel, R. J. (1982): The measurement of anxiety through the computer analysis of verbal samples. *Comprehensive Psychiatry*, 23 (4):364-369.

Gottschalk, L. A. & Bechtel, R. J. (1995): Computerized Measurement of the Content Analysis of Natural Language for Use in Biomedical and Neuropsychiatric Research. *Computer Methods and Programs in Biomedicine, 47:123-130.*

Gottschalk, L. A. & Gleser, G. C. (1969): *The measurement of psychological states through the content analysis of verbal behaviour*. Berkley: California University Press.

Gottschalk, L. A. Winget, C. N. & Gleser, G. C. (1969): *Manual of instructions for using the Gottschalk Gleser content analysis scales: Anxiety, hostility and social alienation - personal disorganization*. Berkeley: University of California Press.

Grünzig, H.-J. (1980): Zur Operationalisierung psychoanalytischer Angstthemen mit Hilfe der computergestützten Inhaltsanalyse (1). In E. Mochmann (Ed.): *Computerstrategien für die Kommunikationsanalyse* (pp. 113-130). Frankfurt: Campus.

Grünzig, H.-J. (1983): Themes of anxiety as psychotherapeutic process variables. In: Minsel, W. R. & Herff, W. (Eds.): *Methodology in psychotherapeutic research. Proceedings of the 1st European conference on psychotherapy research, Trier 1981* (pp. 135-142). Frankfurt: Lang.

Grünzig, H.-J. & Mergenthaler, E. (1986): Computerunterstützte Ansätze. Empirische Untersuchungen am Beispiel der Angstthemen. In: U. Koch & G. Schöfer (Ed.): *Sprachinhaltsanalyse in der psychosomatischen und psychiatrischen Forschung: Grundlagen- und Anwendungsstudien mit den Affektskalen von Gottschalk & Gleser* (pp. 203-212). Weinheim: Psychologie Verlags Union.

Harway, N. I. & Iker, H. P. (1964): Computer analysis of content in psychotherapy. *Psychological Reports*, 14: 720-722.

Henken, V. J. (1976): Banality reinvestigated: A computer-based content analysis of suicidal and forced death documents. *Suicide*, 6 (1): 36-43.

Hogenraad, R. & Bestgen, Y. (1989): On the thread of discourse: homogeneity, trends and rhythms in texts. *Empirical Studies of Arts*, 7:1-22.

Hogenraad, R., McKenzie, D. P., Morval, J. & Ducharme, F. A. (1995): Paper Trials of Psychology: The Words that Made Applied Behavioral Sciences. *Journal of Social Behavior and Personality*, 10:491-516.

Horowitz, M. J., Stinson, C. H., Friedhandler, B., Milbrath, C., Redington, D. J. & Ewert, M. (1993): Pathological grief: An intensive case study. *Psychiatry*, 56: 356-374.

Howe, K. R. (1988): Against the quantitative-qualitative incompatibility thesis or dogmas die hard. *Educational Researcher*, 17: 10-16.

Jeanneau, M. (1991). *Word patterns and psychological structure. Empirical studies of words and expressions related to personality organization.* Umeå: Umeå University.

Johnson, W. (1946): *People in quandaries: The semantic of personal adjustment*. New York; Evanston: Harper & Row.

Kracauer, S. (1972): Für eine qualitative Inhaltsanalyse. *Ästhetik und Kommunikation*, 3:53-58

Krippendorf, K. (1981): *Content analysis. An introduction to its methodology*. Beverly Hills: Sage.

Kucera, H. & Francis, W. N. (1967): *Computational analysis of present-day American English*. Boston: Brown University Press.

Laffal, J. (1968): An approach to the total content analysis of speech in psychotherapy. In J. M. Shlien (Ed.), *Research in psychotherapy* (pp. 277-294). Washington: American Psychological Association.

Luborsky, L., Crits-Christoph, P., Mintz, J. & Auerbach, A. (1988): *Who will benefit from psychotherapy? Predicting therapeutic outcomes*. New York: Basic Books.

Martindale, C. (1978): The therapist-as-fixed-effect fallacy in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 46 (6):1526-1530.

Martindale, C. (1986): Aesthetic evolution. *Poetics*, 15:439-473.

Martindale, C. (1990): *The clockwork muse: the predictability of artistic change*. New York: Basic Books.

Mergenthaler, E. (1985): *Textbank Systems. Computer science applied in the field of psychoanalysis*. Heidelberg: Springer.

Mergenthaler, E. (in press): Emotion-Abstraction Patterns in verbatim protocols: A new way of describing psychotherapeutic processes. *Journal of Consulting and Clinical Psychology.*

Mergenthaler, E. & Bucci, W. (in prep.): *Computer assisted procedures for modeling Referential Activity.*

Mergenthaler, E. & Kächele, H. (1985): Changes of latent meaning structures in psychoanalysis. *Sprache und Datenverarbeitung*, 9:21-28.

Merten, K. (1983): *Inhaltsanalyse - Einführung in Theorie, Methode und Praxis*. Opladen: Westdeutscher Verlag.

Osgood, C. E. (1959): The representational model and relevant research methods. In: I. d. S. Pool (Ed.): *Trends in content analysis* (pp. 33-88). Urbana, IL: University of Illinois Press.

Piaget, J. (1977): *Recherches sur l'abstraction réfléchissante*. Paris: Presses Universitaires de France.

Rosenberg, S. D., Schnurr, P. P. & Oxman, T. E. (1990): Content analysis: A comparison of manual and computerized systems. *Journal of Personality Assessment*, 54:298-310.

Ruoff, A. (1973): *Grundlagen und Methoden der Untersuchung gesprochener Sprache, Bd. 1 (pp. 145, 223-234, 245)*. Tübingen: Niemeyer.

Schöfer, G. (1977): Das Gottschalk-Gleser-Verfahren: Eine Sprachinhaltsanalyse zur Erfassung und Quantifizierung von aggresiven und ängstlichen Elementen. *Zeitschrift für Psychosomatische Medizin und Psychoanalyse*, 23:86-102.

Siemens, R. (1993): Practical content analysis techniques for text-retrieval in large, untagged text-bases. In: P. Beam (Ed.), *Proceedings of the 11th annual international conference* (pp. 293-299). New York: ACM Press.

Speidel, H. (1979): *Entwicklung und Validierung eines Wörterbuches zur maschinell-inhaltsanalytischen Erfassung psychoanalytischer Angstthemen.* Psych. Diplomarbeit, Universität Konstanz.

Spence, D. P. (1970): Human and computer attempts to decode symptom language. *Psychosomatic Medicine*, 32: 615-625.

Stone, P. J., Dunphy, D. C., Smith, M. S. & Ogilvie, D. M. (1966): *The general inquirer: A computer approach to content analysis*. Cambridge: The M.I.T. Press.

Weber, R. P. (1983): Measurement models for content analysis. *Quality and Quantity*, 20: 273-275.

Weber, R. P. (1985): *Basic content analysis*. Beverly Hills: Sage.

Züll, C., Mohler, Ph. P. & Geis, A. (1991): *Computerunterstützte Inhaltsanalyse mit TEXTPACK PC*. Stuttgart: Gustav Fischer Verlag.

Address:

PD Dr. Erhard Mergenthaler, Sektion Informatik in der Psychotherapie, Universität Ulm - Klinikum, Am Hochsträß 8, D-89081 Ulm, Germany, Tel. +49-731/502-5701, Fax: +49-731/502-5662, e-mail: merg@sip.medizin.uni-ulm.de