SSOAR

Open Access Repository
www.ssoar.info

Text analysis and computers

Züll, Cornelia (Ed.); Harkness, Janet (Ed.); Hoffmeyer-Zlotnik, Jürgen H. P. (Ed.)

Veröffentlichungsversion / Published Version
Konferenzband / conference proceedings

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Züll, C., Harkness, J., & Hoffmeyer-Zlotnik, J. H. P. (Eds.). (1996). *Text analysis and computers* (ZUMA-Nachrichten Spezial, 1). Mannheim: Zentrum für Umfragen, Methoden und Analysen -ZUMA-. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-49737-2

Nutzungsbedingungen:
Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.
Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use:
This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.
By using this particular document, you accept the above-stated conditions of use.

gesis Leibniz-Institut für Sozialwissenschaften

Mitglied der Leibniz-Gemeinschaft

Zentrum für Umfragen
Methoden und Analysen

ZUMA-NACHRICHTEN

Spezial

Text Analysis and Computers

May 1996

Cornelia Zuell, Janet Harkness, Juergen H.P. Hoffmeyer-Zlotnik (Eds.)

Copyright © 1996 by ZUMA, Mannheim, Germany

All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

Editors: Cornelia Zuell, Janet Harkness, Juergen H.P. Hoffmeyer-Zlotnik

Contributions to the Text Analysis and Computers Conference, September 18-21, 1995, Mannheim, Germany

Publisher: Zentrum für Umfragen, Methoden und Analysen (ZUMA)
ZUMA is a member of the Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen e. V. (GESIS)

ZUMA Board Chair: Prof. Dr. Max Kaase
Director: Prof. Dr. Peter Ph. Mohler
Post Box 12 21 55
D-68072 Mannheim
Germany

Phone: +49-621-1246-0
Fax: +49-621-1246-100
E-mail cta@zuma-mannheim.de
Internet gopher://gopher.social-science-gesis.de/
http://www.social-science-gesis.de/

Printed by Druck & Kopie Hanel, B4,8, D-68159 Mannheim

ISBN 3-924220-11-5

CONTENT

Foreword 2

Computer-Assisted Content Analysis
Erhard Mergenthaler 3

Computer-Aided Qualitative Data Analysis: An Overview
Udo Kelle 33

Machine-Readable Text Corpora and the Linguistic Description of Languages
Christian Mair 64

Principles of Content Analysis for Information Retrieval Systems
Juergen Krause 76

Conference Abstracts 100

FOREWORD

This ZUMA Nachrichten Spezial documents a cornerstone in an initiative to bring together scholars from different disciplines engaged in the computer-assisted analysis of texts. It presents reprints of the talks given by four keynote speakers and the abstracts of all the papers presented at the Text Analysis and Computers Conference held in Mannheim from September 18 - 21, 1995.

The conference papers were drawn from four broad areas - content analysis, qualitative approaches in the social sciences, information processing and corpus linguistics.

The editors would like to thank once again everyone who contributed to the conference.

One of the main aims of the conference was to provide a forum for an exchange on text analysis procedures and potentials across disciplines - an ambitious undertaking in view of the diversity of perspectives and interests involved.

The conference undoubtedly accomplished some of the ground work necessary for an interdisciplinary discourse to begin. Since September a number of cooperative projects have been started - on new tools for text analysis, on establishing an internet discussion forum and on planning more intensive research cooperation between the humanities and the social sciences.

Clearly, more interchange and, importantly, more cooperation are essential for a single initiative to spread and gain momentum. Together with other institutes, ZUMA is planning further research meetings, an electronic discussion forum and publications. News of developments will be posted regularly on our WWW homepage and in ZUMA publications such as the *ZUMA-Nachrichten*.

May 1996

Cornelia Zuell Janet Harkness Juergen H.-P. Hoffmeyer-Zlotnik

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.

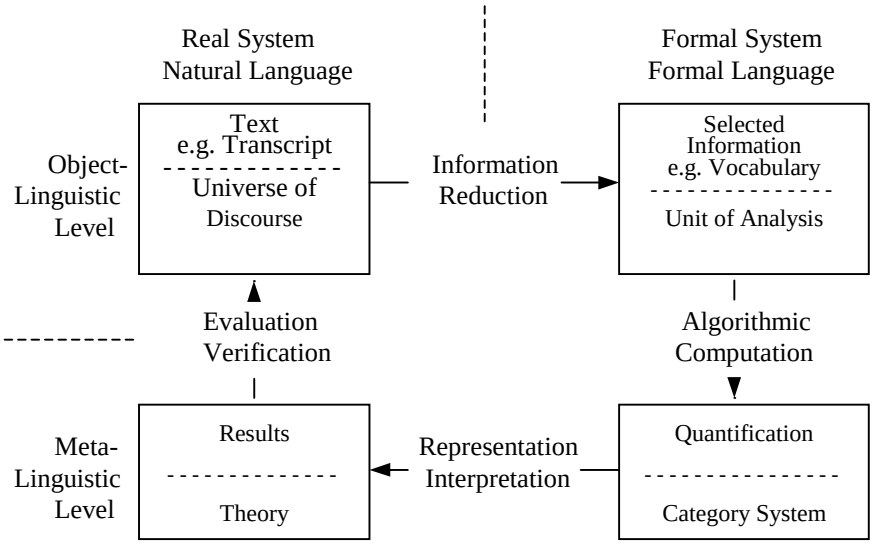

Figure 1: Model of computer-assisted content analysis

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

COMPUTER-AIDED QUALITATIVE DATA ANALYSIS: AN OVERVIEW[1]

Udo Kelle

The last decade saw major advances in computer-assisted coding-and-retrieval methods for non-formatted textual data. These methods are especially helpful in interpretive social research where the researcher has to cope with sometimes huge amounts of unstructured textual data. Presently, qualitative researchers can choose between various coding and retrieval techniques by drawing on a variety of different software packages. The purpose of the paper is to give an overview of computer-aided techniques for the management and analysis of textual data in qualitative research and of the current debate about the methodological impact of these techniques on the research process. The initial sections contain a brief historical overview of the development of computer-aided qualitative data analysis whereby some epistemological aspects of the relationship between qualitative methodology and computer-use will also be discussed. Following that the paper also outlines basic elements of "computer-aided qualitative data analysis", namely the use of textual database management systems for the automatization of manual indexing and operations. Since the advent of the first coding-and-retrieve programs great hopes have been expressed that such coding techniques could revolutionize qualitative research by making the research process more transparent and by improving the reliability and validity of its results. In the last part of the paper these questions will be discussed thereby focussing on aspects of validity.

1. Introduction

In the past decade a variety of software programs have been developed to assist qualitative researchers[2] in analysing their data. More and more researchers now use these programs and there is a growing body of technical as well as sophisticated methodological literature about computer-aided qualitative data analysis (cf. LeCompte & Preissle, 1993: 279-314; Lee & Fielding, 1991; Richards & Richards, 1991; Tesch, 1990; Kelle, 1995; Weitzman & Miles, 1995). As with other technical innovations in their early stages one

can find enthusiastic forecasts and concerned warnings about their possible merits and dangers. While some qualitative researchers warn that computer-aided methods might alienate the researcher from their data (cf. Agar, 1991, Seidel, 1991), others are really thrilled by the prospect that computers could add trustworthiness to qualitative inquiry and foresee a methodological revolution (Richards & Richards, 1991).

In the following overview of the current state of the art of computer-use in interpretative research I will try to present some preliminary answers to the question of its potential methodological costs and benefits. I will start by giving an introduction to the basics of computer-aided qualitative data administration, in particular the techniques of coding and retrieval. The first programs specifically developed for managing qualitative data in the early eighties were based on these techniques and facilitated the mechanization of rather mundane mechanical tasks, namely the building of indexes, concordances and index card systems. These programs, e.g. Qualpro, The Ethnograph or Hyperqual are sometimes referred to as the *second generation* of computer-aided qualitative data analysis, while the first generation were word-processors and standard database management systems (Mangabeira, 1995:130). Second generation programs, in particular The Ethnograph, are now widely spread within the qualitative research community and it is now possible to draw on a growing body of practical experience when discussing their methodological impact on qualitative research.

The situation is completely different if one looks at the *third generation* of programs for analysis which (although they are based on the same principles as the second generation software) contain a variety of features that greatly exceed manual methods of textual data administration, for example Atlas/ti, HyperResearch, Aquad or NUD•IST. These programs have now been on the market for some years but the extended features are only seldom used as recent investigations among qualitative researchers have demonstrated (Dotzler, 1995; Lee, 1995). It is not yet clear whether this is due to a certain technological conservatism that adherents of the qualitative paradigm are supposed to share or whether the advanced features of these programs are not really useful for the purposes of qualitative research.

In the last part of my paper I will advocate the latter position by arguing that some of the extended features provided by third generation programs no longer support qualitative, interpretive analysis but require a style of coding of qualitative data that is much closer to that applied in classical content analysis.

2. Computer-aided Methods for the Management of Textual Data

Let me start with some short remarks about the history of computer use in qualitative research. Programs for the statistical analysis of textual data have been available since the mid-sixties. In 1966 *The GENERAL INQUIRER*, a program for computerized quantitative content analysis, began a train of development in history, linguistics and literary studies that led to the emergence of a whole scientific community concerned with computing in the humanities. In the social sciences, however, the use of software for computer-aided textual analysis initially only attracted scholars working in the field of content analysis. Qualitative researchers from interpretive traditions such as Chicago School sociology or ethnography who also used texts as their main (if not only) empirical data source made no attempts to integrate such software into their analytic work.

This is not at all surprising if one takes into account that qualitative analysis in these traditions meant a totally different style of analytic work than that found in content analysis. For the interpretive traditions textual analysis usually consists of a thorough, fine grained analysis of a text in order grasp its meaning through hermeneutic understanding - an operation that is often viewed as an artistic endeavour ("*hermeneutische Kunstlehre*"). Quantitative content analysis was criticized by such scholars as being too atomistic and oversimplistic to really capture the semantic content of texts. In contrast, hermeneutic analysis was considered to be the method that was capable of taking into account the ambiguity and context-relatedness which were regarded as the central characteristics of everyday language use (cf. Giddens, 1976).

The opinion that computers were not at all useful for textual analysis was supported by the paradigm of computer-use prevalent in the era of the mainframe. Computers were mainly seen as calculating machines; useful in the social sciences only for statistical analysis. The idea that electronic data processing machines could one day become an indispensable tool for the storage, retrieval and manipulation of text and thus also helpful to qualitative researchers was far away.

This situation was radically changed by the advent of the Personal Computer. In the mid 1980s many *hommes des lettres,* qualitative researchers among them, discovered rather quickly the enormous possibilities for text manipulation that were offered by the new technology. But, given the limited user-friendliness of early operating systems and software environments, many users (especially those working in a DOS environment) were also compelled to acquire a certain expertise in computer-use. After a strenuous apprenticeship, many of them experienced real enthusiasm when they discovered the numerous

possibilities for working with textual data offered by the new technology. Consequently, the dominant paradigm of computer-use changed from "computers as number-crunchers" to computers as devices for the intelligent management of data, incorporating facilities for the storage and retrieval of information that were far more complex and convenient than any manual system of information retrieval used previously.

With this paradigm shift it became clear that, although computers are not useful for the hermeneutic analysis of text, they can nevertheless be of great assistance to a hermeneutician. Since hermeneutic analysis tends to produce huge amounts of unstructured textual data, such as interview transcripts, protocols, field notes, and personal documents, there are many data storage and retrieval tasks involved in this kind of analysis. In hermeneutic sciences various strategies of intellectual craftsmanship have been developed to manage these tasks and to keep track of one's data. Many of these techniques are several hundred years old and widely used in all sciences that work with texts - in fact most of them were already used in the context of biblical exegesis.

1. *Building indexes*: on a separate piece of paper, the researcher notes the place (in terms of line, page, interview number) where a certain subject is discussed by the interviewee. The result of this process is similar to the index in a book.

Name Index

Agar, 9, 12, 60	Barton, 135	Blackman, 28(N)
Aldenderfer, 165	Baszanger, 50(N)	Bogdan, 5, 57, 61
Altheide, 20	Becker, 4, 56	Bradshaw, 180
Anderberg, 165	Berelson, 53	Brownstein, 51(N)
Araujo, 9, 13, 68	Bertaux, 50(N)	Bryman, 152
Bailey, 165	Biklen 5, 57, 61	Burgess, 153
Bain, 22	Blashfield, 165	Campbell, 22, 152

Figure 1: Name Index

2. Including *cross references* in texts telling the reader where to find more information on the same subject, for example as seen in the margins of a Bible.

3. *Decontextualization* and *comparison of text passages*: Before the advent of computers, "cut-and-paste" techniques were the most widely used methods of organizing the data material to facilitate the comparisons of text passages - the researcher had to "cut up field notes, transcripts and other materials and place data relating to each coding category in a

separate file folder or manila envelope" (Taylor and Bogdan, 1984:136; see also Lofland & Lofland, 1984: 134). Other researchers used index cards for this purpose.

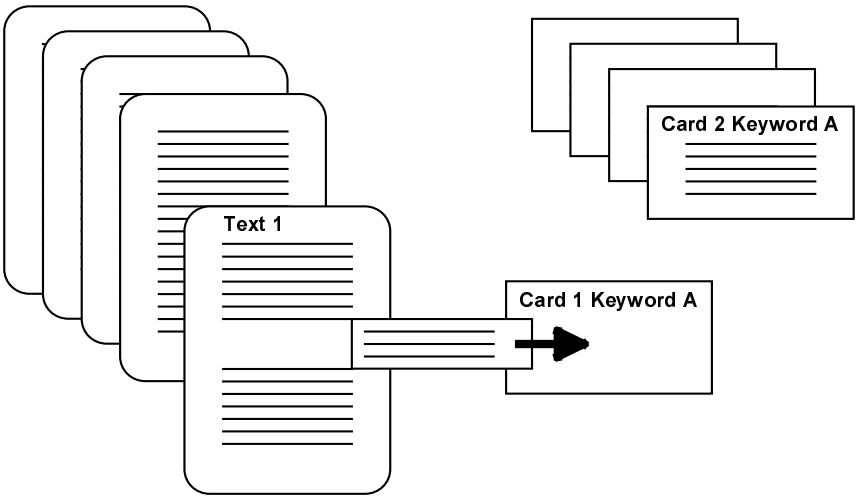

Figure 2: Cut and paste

Unfortunately such techniques of data management are rarely made explicit (with some exceptions, e.g. Miles & Huberman's source book (1994), Tesch (1990)). Instead they form a part of the "folklore" of qualitative research, i.e. a departmental tradition passed on orally among colleagues. From a methodological perspective such techniques are far from trivial, on the contrary, their methodological significance for hermeneutic research in the humanities or *Geisteswissenschaften* can hardly be overrated.[3]

The same holds true for qualitative studies in which a large amount of textual data are collected, for instance when 30 to 40 interviews or more are conducted. In such cases the validity of the study's results is highly dependent on the "folkloristic techniques" applied: A researcher who had organized his/her material in a sloppy way will certainly overlook and neglect crucial information, and his/her inferences and conclusions will be flawed because they are based on sparse data material and counter-evidence has not been systematically considered.

Although they are of great methodological significance, the manual methods of data management also have certain serious constraints. Firstly, they are rather inflexible. The code scheme cannot be modified easily, a restriction that contradicts the fundamentals of an inductive style of analysis which is typical for qualitative research. This inductive style requires that the category scheme is developed from the data and not constructed beforehand and then imposed on the data. Furthermore, if one uses "cut-and-paste" techniques to decontextualize text segments it is almost impossible to later enlarge an extracted text segment. In many studies this has become a major problem, especially since novices tend to choose segments that are too small. It can happen that when, several weeks later, the researcher returns to re-read the text passages that they have cut out and pasted on index cards, the decontextualized text segments turn out to be totally unintelligible. One strategy for coping with this problem is to include with the text passage information about its original site, so that the researcher could trace the path back to its context (Miles & Huberman, 1984:106). However: "*Cards and file folders are reasonably workable if the number of sites is small and the data collection not extended. But they are increasingly difficult and very time-consuming as the database gets larger"* (Miles & Huberman, 1984:67). The construction of an index will bring about the same disadvantage which tends to negate the advantage that it offers by leaving the text passages in their original context. As the database grows the search for text passages and especially their comparison becomes an increasingly tedious task.

This problem of "data overload" (Miles & Huberman, 1994) is often mentioned in the technical literature about qualitative analysis, and it is aggravated by a second problem. Since in interpretive analysis data analysis and theory construction are closely interlinked, the researcher generates many theoretical concepts in the process of data analysis which are often recorded as memos across numerous notebooks, manuscript pages and index cards. A central step in qualitative analysis is to compare the different text segments and memos in order to identify commonalities, differences or linkages between them. The purpose of this is to identify structures and to construct "meaningful patterns of facts" as Jorgenson (1989:107) put it. The crucial problem in the hermeneutic analysis of large amounts of textual data is that at any given point the analyst must be able to draw together all text passages, chunks of data and memos that relate to a certain topic. If one considers that an average study's database consists of 30 interviews with a transcribed length of around 30 to 40 pages one can easily imagine that this can be a mammoth organizational task.

It was the search for solutions to these problems that led to the first attempts to computerize cut-and-paste methods in the early 1980s. When the first qualitative researchers came to realize the advantages of word processing systems for writing texts, they tried to use these programs for managing their textual data. Manual cut-and-paste techniques were computerized by copying text segments from one file to another. As one can easily see, this method has hardly any advantage over manual methods - it simply replaces scissors, paper and glue with an electronic cut-and-paste facility. Other researchers started to experiment with the database programs available for microcomputers. But, although these programs permitted the storage and retrieval of text segments according to as many criteria as necessary, and also offered search and sorting procedures, they still imposed serious limitations on the management of unstructured textual data. For example, standard database management systems such as dBase require that the text segments are stored in a field that is defined before the data are entered. As has been mentioned above, this contravenes the inductive categorization strategy preferred by most qualitative researchers.

Inspite of these drawbacks, some researchers used the macro or programming language often contained in standard software like word processors and database management systems to adapt these programs to their specific requirements. The results of these endeavours can be regarded as the *first generation* of programs for computer-aided textual data management in qualitative research. But it did not take long before some qualitative researchers with advanced computer programming skills started to develop *non-formatted textual database systems* for the management of unstructured textual data. The idea behind this kind of database management systems is straightforward: the addresses of certain text segments, in terms of line numbers, are stored as pointers in a special file together with the names of the codes allocated to these segments.

No of Document	Name of Code	First line	Last line
1	CLE	234	245
1	EMO	167	201
2	CLE	56	88
2	CLE	195	209
2	EMO	355	390

Figure 3: Codes as pointers

These pointers can be used by retrieval algorithms to find and retrieve those text segments to which certain codes were assigned by the researcher. Therefore, as opposed to standard database systems or word processors, coding with these programs does not entail the removal of text segments from their context. Text segments are only temporarily decontextualized for retrieval purposes. Consequently, the use of non-formatted textual database system has two central virtues: (1) in principle it is possible to electronically restore the original context of a text segment, and (2) the coding scheme can be changed much more easily than with a standard database.

Non-formatted textual database management systems represented the second generation of software for qualitative research. Since the mid-1980s a number of different software packages, like Qualpro, The Ethnograph, Textbase Alpha or Max have been developed that are based on these principles. All these programs allowed the researcher to code their data, that is to attach codes to certain text segments and to retrieve all text segments from a defined set of documents to which the same code had been assigned (see Figure 4).

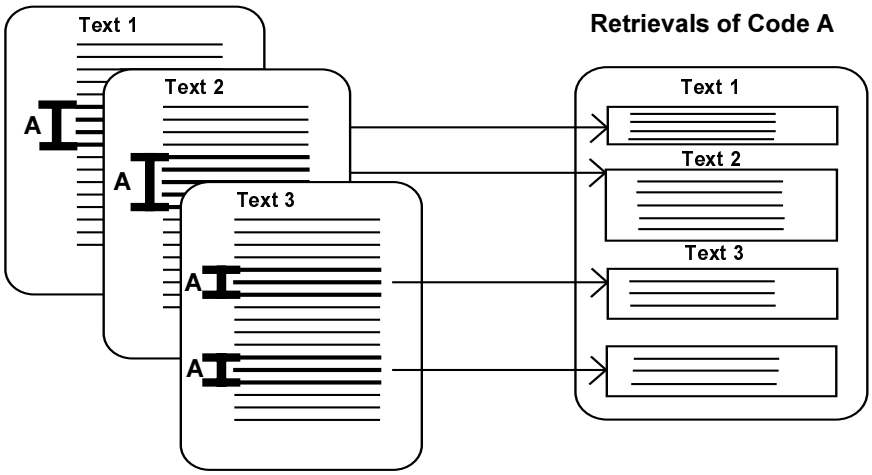

Figure 4: Coding and retrieval

These programs vary greatly with respect to their user-friendliness and they contain different additional features, e.g. facilities for editing and storing lengthy theoretical

comments or statistical features for calculating code or word frequencies. But since their basic function - the coding and retrieval of text segments - is the same these software packages can be subsumed under the heading "code-and-retrieve" programs.

The integration of code-and-retrieve programs into the research process can be regarded as a major methodological innovation. To understand why this is the case it is useful to take a closer look at the logic of qualitative reasoning. In analysing qualitative data the analyst usually does not start with ready-made hypotheses but with a broad and general heuristic framework of theories. While exploring, that is reading, re-reading and interpreting the data, the researcher will develop assumptions about associations, regularities and patterns in the field under investigation. Further analysis means the modification and concretization of these assumptions. This process often also contains elements of "proving" and "checking" them, and qualitative methodologists have often referred to this process as "examination" or even "verification" of "hypotheses" (cf. Strauss, 1987:12, Miles & Huberman, 1994:262, Strauss & Corbin, 1990:108). But in using these terms one must be very careful to not obscure the differences between the assumptions that a qualitative researcher develops in the ongoing process of analysis on the one hand and statistical hypotheses on the other. Qualitative hypotheses, when they first come into a researcher's mind, are usually not highly specified and definite propositions about certain facts, but tentative and imprecise, sometimes vague conjectures about possible relationships. Following the philosopher of science Norwood Hanson one should, instead of calling them hypotheses, rather call them hypotheses about what *kind* of propositions descriptions or explanations will be useful in the further analysis. They are insights that "*whatever specific claim the successful H(ypothesis) will make, it will nonetheless be an hypothesis of one kind rather than another.*" (Hanson, 1971:291) A researcher who conducts a field study about the distribution of power in a certain organization may, for example, gain the impression from the first interview that there is some sort of hidden competition between the different divisions of the organization. Or a researcher who carries out qualitative interviews among patients suffering from chronic pain may for example initially develop the idea that women form different pain management strategies to men. Using further empirical material the qualitative researcher would now try to further elaborate these tentative conjectures. As the hypotheses become more elaborated, they will also become more precise, gain empirical content and thus will come closer to hypotheses in the original sense, that means to empirically testable statements about distinct entities.

Only at this point in the ongoing process of analysis will it be useful to talk about "verification" or "hypothesis testing". However from the preceding discussion it should now be obvious that these terms have to be used in a completely different way than in

quantitative research. But, in the contemporary methodological literature about qualitative research the authors hardly ever make explicit what they understand by "hypothesis examination", "verification" or "falsification". One has to go back to the tradition of the late Chicago School in the 1950s to find an explicit account of qualitative hypothesis examination.

Lindesmith and Cressey (Lindesmith, 1968; Cressey, 1950; 1971) in seeking to apply the strategy of analytic induction outlined by Znaniecki (1934) to research practice proposed a methodology of hypothesis examination in which the researcher starts by formulating a vague definition together with a hypothetical explanation of the investigated phenomenon. Thereafter a single case is examined in the light of the hypothetical explanation to determine whether the hypothesis can account for the investigated phenomenon in this case. If not, either the hypothesis has to be reformulated or the investigated phenomenon has to be redefined in such a way that the case can be excluded - of course the new definition has to be more precise than the preceding one in order to avoid the immunization of the hypothetical explanation. "...this procedure of examining cases, re-defining the phenomenon and re-formulating the hypothesis is continued until a universal relationship is established, each negative case calling for a re-definition or re-formulation" (Cressey, 1971: 16).

Lindesmith regarded this methodology of qualitative hypothesis examination as a fallibilistic strategy in the Popperian tradition of critical realism claiming that by looking for crucial cases the researcher systematically exposes their hypothesis to the possibility of failure. Nevertheless, this strategy differs greatly from the concept of statistical hypothesis testing, because

1. firstly, it cannot be regarded as the application of a set of precisely defined rules. Furthermore, the guidelines (not "rules") outlined by Lindesmith and Cressey can be seen as a heuristic framework that help researchers to develop a theory via the successive refinement of working hypotheses.

2. Secondly, the empirical material not only serves as the basis for making a decision about the rejection or acceptance of a hypothesis, but also as an information source for the generation, refinement and modification of new and existing hypotheses.

In this context, "testing and confirming findings" or "verification" means: returning to the data (i.e. re-reading one’s transcripts or field notes), or returning to the field (i.e. conducting new observations or interviews), in order to find some confirming or disconfirming evidence.

However, some serious threats for validity are associated with this strategy of hypothesis examination. The researcher is always in danger of treating their material selectively, that is of only noticing confirming evidence and overlooking disconfirming evidence. This potential danger is, as one can easily imagine, aggravated by data overload. A integral part of the folklore of qualitative research is anecdotal accounts of researchers who have thousands of pages of transcripts available and who desperately try to find at least two or three text passages that support their assumptions which they can quote in a publication.

It is obvious that this danger diminishes if the data material is well organized within a coding and retrieval system that allows the researcher to easily draw together all text passages that refer to a certain topic. Thus computer-aided methods for the administration of textual data will help to fully exploit the data material and prevent researchers from basing their results on sparse evidence. By reducing the negative effects of data overload they will also allow researchers to collect and analyse more textual data. As with hypothesis testing this should not seduce us into confusing different modes of sampling: the purpose of qualitative sampling cannot be to achieve representative samples in the statistical sense since this would require far more cases then could be analysed by means of interpretive methods. But methods of purposeful sampling regularly applied in qualitative research, for example "theoretical sampling" (Glaser & Strauss, 1967:45ff), can also benefit from greater sample sizes: such a strategy can be used to systematically search for crucial cases, i.e. cases with a high probability for providing evidence or counter-evidence for the developing hypotheses.

From this it can be concluded that second generation software represents a major methodological innovation for qualitative research since they allowed the researcher to analyse more data more systematically and carefully and thus increased the possibility to find evidence or counter-evidence for their hypotheses.

3. Enhanced Coding and Retrieval Techniques

Code-and-retrieve programs only mechanized widely used cut-and-paste or indexing techniques but did not change their underlying logic or offer analytic features which could not be employed using manual methods. This situation changed as more and more features were added to these programs, features that widely exceeded the analytic possibilities offered by manual methods, e.g.

1. facilities for *theory building* that offered the researcher the possibility of defining many kinds of linkages between codes, memos and text segments resulting in complex networks

2. tools for *hypothesis examination* that are based on complex retrieval techniques for searching for co-occurring codes.

These new techniques formed the basis of what have been called the *third generation programs* (Mangabeira, 1995). In the following I will briefly outline some of their basic principles and then move on to discuss their methodological impact on the research process. At present literature about the methodological advantages or disadvantages of the advanced theory-building and hypothesis testing capabilities of third generation programs are mainly published by the developers themselves. In contrast, very little has been written by neutral researchers about the practical experience of using these techniques. Therefore, I will draw on some general considerations about the nature of qualitative research to evaluate their possible methodological costs and benefits.

3.1 Enhanced Coding Techniques for Constructing Linkages and Building Networks

The theoretical categories which are developed in the ongoing process of qualitative analysis are often closely related to the codes used for coding the text. Since one would usually regard theories as networks of theoretical categories it is obvious that tools for connecting codes to each other could be helpful for displaying the structure of the emerging theory. Therefore, software that facilitates the connection of categories can make a major contribution towards theory building. As has been already mentioned, coding is technically realized in most code-and-retrieve programs by defining pointers. These pointers contain the addresses of text segments and thus establish linkages between codes and text segments. In the same way it is also possible to define a linkage between one code and another. This linkage can take the form of, for example, the subsumption of one code under a more general code, or the subdivision of one code into several more refined subcategories. If the researcher restricts themselves to this kind of connection their category scheme could be represented by hierarchical networks (see figure 5).

The program NUD•IST contains extensive features which support the construction of hierarchies of code categories. But linkages between codes may not only take the form of hierarchical relations but can form whole networks of categories, containing chains or loops (see Figure 6.).

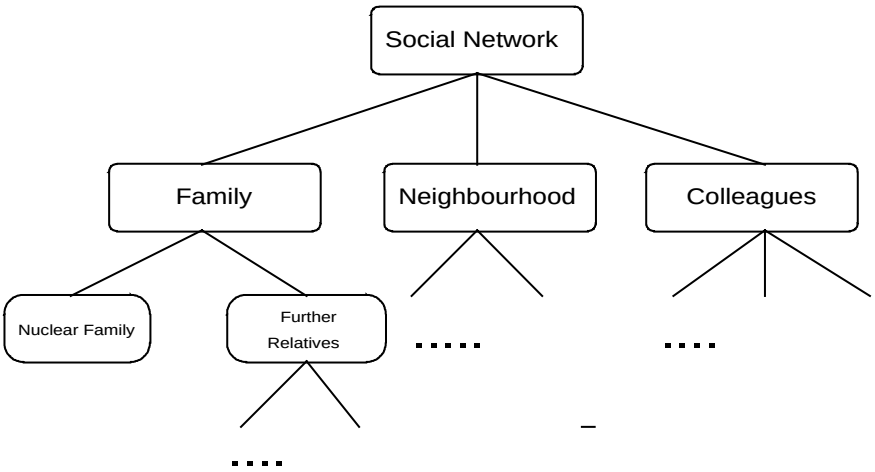

Figure 5: hierarchical network of code categories

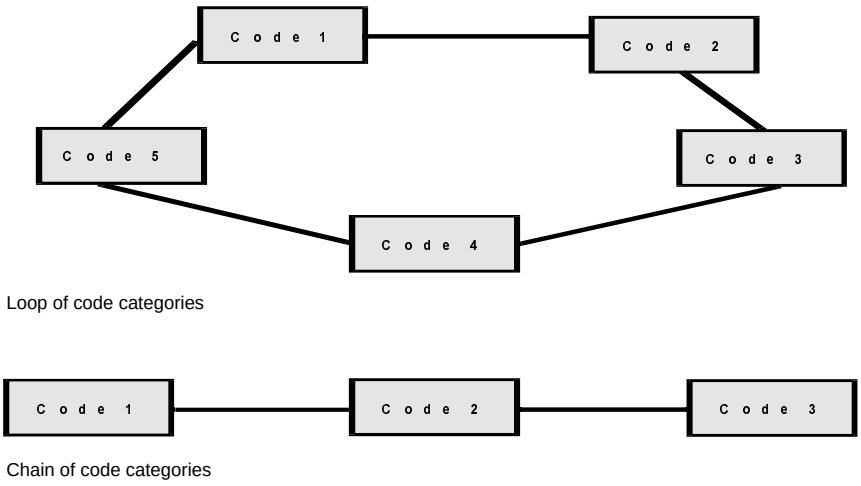

Figure 6: different kinds of networks of code categories

The program Atlas/ti offers a variety of features for building non-hierarchical networks. Additionally the user of this program can label the linkages between codes as being, for example, causal, contradictory, or hierarchical relationships, while Hypersoft, which is another program with extended theory building facilities, offers tools for defining the strength of a relation in quantitative terms.

It is also possible to link "memos" (see above) to other elements of the database. Since memos are often *"the theorizing write-up of ideas about codes and their relationships as they strike the analyst while coding"*, as Glaser (1978: 83) said, they are most useful if linked to the relevant codes or text segments. Since memos can be regarded as signposts along the path between data and theory, linkages between text segments, codes and memos can help to retrace this path, enabling researchers to control the empirical background of their theoretical ideas. Further possibilities for linking elements of the qualitative database include the employment of *hyperlinks* for linking text segments to each other without using codes. In this way the technique of defining cross-references already mentioned earlier can be mechanized.

3.2 Enhanced Retrieval Techniques for Searching for Co-occurring Codes

In some of the third generation software programs, for example NUD•IST, these enhanced coding features for theory building are supplemented by enhanced retrieval techniques that are intended to help with the examination of hypotheses and thus support the "grounding" (in the sense Glaser and Strauss (1967) used this term) of the emerging theories in the data.

The first *enhanced retrieval facilities* were already added to second generation programs in the mid-80s. The researcher could define case-constant variables, such as age, gender, profession, assign them to the cases and then *selectively retrieve* text segments according to two criteria: First, that they have been coded with a certain code and second, that they belong to a specified subgroup of documents defined according a certain value of a case-constant variable. For example a researcher could examine the hypothesis that men and women generally have different attitudes towards a certain topic by searching for all statements on this topic (i.e. a certain code) made by male interviewees (i.e. value "male" of the case-constant variable "gender") and comparing them with those statements made by women. Or utterances made by respondents from different age groups or from different professions could be compared.

The next step in the development of enhanced retrieval facilities were algorithms for searching for *co-occurring codes*.

Facilities for searching for co-occurring codes can be regarded as a possible basis of a methodology of hypothesis testing in qualitative research (see Sibert & Shelly, 1995, Hesse-Biber & Dupuis, 1995, Huber, 1995). This is usually technically realized with methods of logic programming. The first step is the construction of a knowledge base that contains information about what codes are connected to which parts of the textual database. The researcher then formulates their hypotheses as relationships between code categories which, finally, are broken down in such a way that a query to the knowledge base can be conducted. This query is designed to provide information on whether the categories in question co-occur in the text

HyperResearch is one program that implements such ideas. As with second generation programs HyperResearch's basic principle is coding and retrieving text segments. In addition the program contains a hypothesis testing module which is designed to formalize the inference or thought process of a qualitative researcher analysing texts by inferring new codes from existing codes. When using the hypothesis testing module the researcher formulates their hypotheses in the form of "production rules" in which codes are connected with "if-then" statements. To give an example: a researcher who has coded his/her data with codes for "critical life events" and "emotional disturbances" may wish to examine the hypothesis that critical life events are always or frequently accompanied by emotional disturbances. They could then transform their hypothesis into a query about all co-occurrences of text segments coded as critical life event with segments coded as emotional disturbance. Using HyperResearch's hypothesis tester one would formulate the rule

IF "critical life events" AND "emotional disturbances" THEN ADD "life event has caused stress".

If the program finds both the code "critical life events" and the code "emotional disturbances" in a given document, the hypothesis is confirmed for that document and the code "life event has caused stress" is added to it.

HyperResearch only searches for the presence of certain codes within a given set of documents, and in doing so does not take the precise location of the text segments into account. In contrast, with programs like Aquad, NUD•IST or Qualog co-occurrences of codes can be defined more precisely. They can be:

- Indicated by *overlapping* or *nested* text segments to which the codes under investigation are attached, as shown in Figure 7.

- Indicated by text segments that are coded with certain codes (here A and B) that appear with a certain *specified maximum distance* of each other. If this maximum distance is set at, say, 8 lines, the program would retrieve all code A. (see Figure 8)
- indicated by *sequential ordering* (Code A is regularly followed by Code B), as shown in Figure 8.

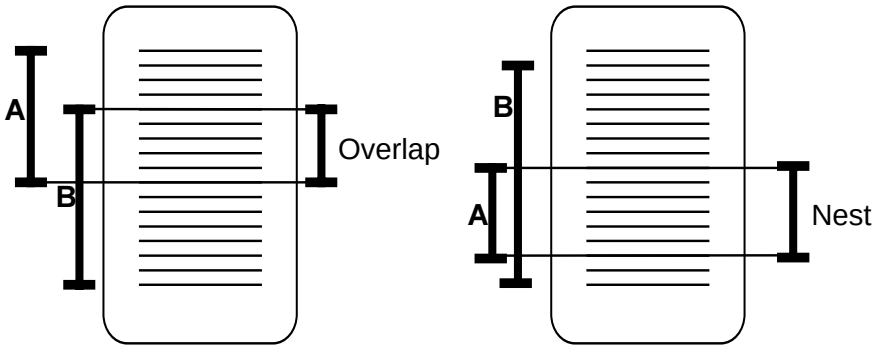

Figure 7: Overlapping and nesting text segments

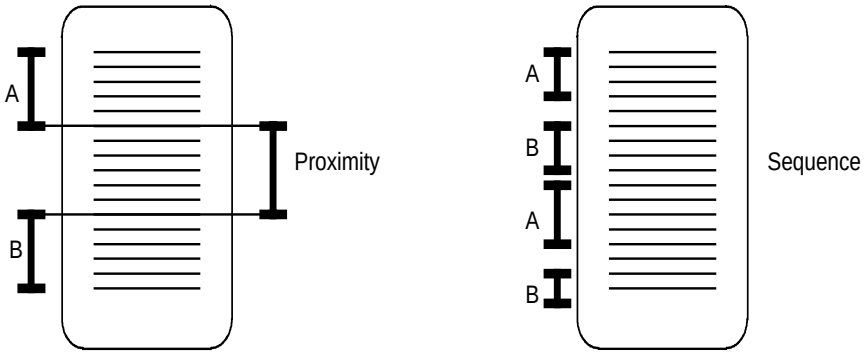

Figure 8: Proximity and sequence of codes

The program Aquad is an example of a program that helps the researcher to use information about the co-occurrence of codes for hypotheses examination. Aquad contains twelve ready-made hypothesis structures that are formulated as searches for co-occurring codes. These hypothesis structures can be expanded by users who are familiar with the Prolog programming language in which the whole program has been written. Taking our previous example and using Aquad one would first code the text segments with the codes "*cle*" (for "critical life events") and "*emo*" (for "emotional disturbances"). Let us assume that during this process the following hypothesis has come to the researcher´s mind "Whenever interviewees talk about critical life events they will also, at the same time, mention emotional disturbances". One can now operationalize "at the same time" as "within a maximum distance of 5 lines in the interview transcript" and run a retrieval that finds all text segments coded with "*cle*" where a text segment coded with "*emo*" also occurs within a maximum distance of 5 lines. Looking at the result of such a retrieval shown in Figure 9 one can for example see that in the interview "bioss1" the association of "*cle*" and "*emo*" occurs only once (at line 102), while in interview "bioss2" there are 5 text passages where text segments coded with these codes are very close to each other.

Thus the co-occurrence of codes (defined as the overlapping, nesting, proximity or sequential ordering of text segments) indicates the presence of critical evidence for or against the hypothesis.

hypothesis 1 / codefile **bioss1.cod**

100	102	**cle**	**-**	102	104	**emo**

hypothesis 1/ codefile **bioss2.cod**

24	28	**cle**	**-**	26	30	**emo**
65	70	**cle**	**-**	72	82	**emo**
110	112	**cle**	**-**	111	115	**emo**
220	228	**cle**	**-**	212	224	**emo**
450	452	**cle**	**-**	456	476	**emo**

Figure 9: result of a co-occurring code search with Aquad

A similar technique called "Qualitative Comparative Analysis (QCA)" (Ragin, 1987, 1995) that uses enhanced retrieval techniques can be realized by both Aquad and the

program QCA. This method is similar to the hypothesis testing module contained in HyperResearch in that the presence or absence of codes in a certain case is the decisive criterion. However, with QCA it is possible to investigate far more complex configurations of codes. Ragin maintains that the application of this approach is especially useful for researchers who wish to investigate complex causal relationships. For this the material has to be coded in such a way that codes represent possible conditions, causes and outcomes. It should be pointed out that what are called "possible causes" and "outcomes" are solely defined by the user - as with HyperResearch QCA can be regarded as a tool for formalizing the thought processes of a qualitative researcher and not as a tool for "proving" hypothesized causal relationships.

QCA uses Boolean Algebra to determine the subset of codes (from a set of codes defined as "conditions" or "causes") which is essential for the occurrence of the code defined as "outcome". For this purpose, a strategy of redundancy elimination is applied, using the logical minimization of truth tables.

Let me give a simple example to illustrate this process. Let us define the two codes "*male*" (M) and "*parents' high socio-economic status*" (P) as causes and "*respondent's high socio-economic status*" (S) as the outcome. There are eight possible configurations of these three codes, as indicated in Figure 10.

	Causes		Outcome	
	M	P	S	
1.	0	0	1	
2.	0	0	0	x
3.	0	1	1	
4.	0	1	0	
5.	1	0	0	
6.	1	0	1	
7.	1	1	0	
8.	1	1	1	x

Figure 10: Example of Boolean Minimization

If only configurations 2.) and 8.) are actually present in the data then it is concluded that the outcome S is only achieved if both M and P are present. If additionally configurations 3.) and 6.) are realized it is concluded that the outcome S can also be achieved if only one of the causes M or P is present. Although this is a simple example the method can be extended to the investigation of numerous causes.

4. The Methodological Relevance of Enhanced Retrieval Techniques

It is interesting to note that recent investigations among users of qualitative analysis software have clearly shown that enhanced retrieval techniques are only seldom used (cf. Dotzler, 1995; Lee, 1995). It is not yet known whether this is due to the reluctance of users to adopt new analytic techniques or whether these techniques are inadequate tools in the context of qualitative research.

Due to the lack of practical experience with these techniques at the moment it will be only possible to give an answer to this question on the basis of general theoretical and methodological reflections. Since enhanced retrieval techniques are regarded by some authors as the basis of a methodology of qualitative hypothesis testing it will be useful to clarify first of all the different concepts of a hypothesis and hypothesis testing in the context of quantitative and qualitative research.

However, this attempt of clarification will be complicated by the fact that the notions of a hypothesis and hypothesis testing are themselves controversial among qualitative researchers and methodologists. On the one hand many authors often emphasize that qualitative analysis contains elements of hypothesis examination and verification (e.g. Miles & Huberman, 1994: 262ff; Strauss, 1987:11ff; Strauss and Corbin, 1990:107ff) From this perspective qualitative analysis is a series of alternating inductive and deductive steps: data-driven inductive hypothesis generation is followed by deductive hypothesis examination for the purpose of "validation", or "verification".

On the other hand scholars devoted to a relativistic epistemology rooted in postmodernist and constructivist philosophy would strongly object to the idea that qualitative research has anything to do with "verification" or the "testing of hypotheses" - after all, the concepts of hypothesis examination and verification form an integral part of standard social science methodologies which have been criticized by many qualitative researchers for imposing methodological models from the natural sciences onto social research (see e.g. Denzin & Lincoln, 1994: 101).

I think one will benefit from looking at the usage of the term "hypothesis testing" in its traditional domain of statistical analysis. Statistical hypothesis testing (or "significance testing") is a strictly rule-governed process of assessing the statistical significance of empirical results, whereby empirically observed sample findings are compared with theoretical expectations. This comparison requires the computation of the probability that the observed outcome is the result of mere chance. On the basis of this calculation a precisely formulated decision rule can be applied: if the probability that a certain result is merely the effect of a random process is above the so-called alpha-level, then the researcher has to reject the hypothesis.

4.1 Hypothesis Testing in Quantitative Content Analysis

The application of such a methodology to the analysis of textual data has certain important prerequisites which can be best explicated by taking quantitative content analysis as an example. Quantitative content analysis parallels qualitative research in certain ways: unstructured textual data are also often used as the primary data source and coding of these data is also the first step of data analysis. But coding in quantitative content analysis serves a quite different purpose than qualitative coding: codes do not primarily have an index function to help identify text passages relating to a certain topic. Instead, codes in quantitative content analysis represent values of certain variables. Each appearance of a certain code represents a certain event that is of interest to the researcher. For example, coding a text with the code "Liberal Party affiliation" would mean that the interviewee is a supporter of the Liberal Party. Coding is usually followed by information reduction which entails using the information provided by coding for the construction of a new (quantitative) data corpus that can be analysed with statistical procedures. The frequencies of certain codes can be calculated, and hypotheses about the co-occurrence of codes can be tested. Unlike qualitative research, it is crucial for this kind of analysis to focus almost exclusively on the codes and not on the raw data. But this is only possible if the codes can be seen as true representations of certain facts described by the raw data. Consequently, there has to be a high degree of certainty that the codes have been applied in a systematic and consistent way, in other words, the coding must have a high degree of validity and reliability. Furthermore, the coding of the raw data must be inclusive and exhaustive. This means that one must be certain that every single instance of the investigated phenomenon that occurs in the raw data has been coded.

These requirements make it essential that whenever codes are employed to condense the information contained in the data by representing facts described by the data, a *precise coding scheme* is developed ***before*** coding starts, since:

1. For pragmatic reasons alone, inclusive and exhaustive coding would not be possible if the researcher did not have a ready-made category scheme to hand right from the start. If, instead, the coding scheme was being permanently altered, it would be necessary to permanently re-code the previously coded data with the newly developed categories.

2. Objective, and therefore reliable, coding can only be conducted if all coders employ exactly the same coding scheme.

Consequently, a research strategy where codes are used to test hypotheses in the sense outlined above requires a deductive approach: the relevant variables and their values (that form the codes) have to be determined before data are coded.

4.2 Hypothesis Testing in Qualitative Research

However, this requirement for a deductive approach would cause severe difficulties in the context of interpretive social research.

Let us again take a short look at the logic of qualitative reasoning to clarify this point. Although there is a puzzling heterogeneity of qualitative approaches (cf. Tesch, 1990: 55ff.), it has often been emphasized that most of them are at least implicitly based on a common underlying concept of human action. This has been referred to with notions like the "*Interpretive Paradigm*" (Wilson, 1970), "*Interpretive Sociology*" (Giddens, 1976) or "*Interpretive Interactionism*" (Denzin, 1989).

According to the interpretive paradigm the meaning of human action and interaction can only be adequately understood if the interpretations and the common-sense knowledge of the actors are taken into account. This theoretical postulate has far-reaching methodological consequences: The researcher must be able to gain access to the interpretations and the common-sense knowledge of the members of the social world investigated. If the researcher's goal is to describe members' actions adequately (which is with respect to the meaning these actions have for them) he/she must be able, to a certain extent, to perceive the world in the same way as the members do.

This demand for "empathetic understanding" of or access to the common-sense knowledge of the investigated form of social life makes it difficult, if not impossible, to employ a hypothetico-deductive (H-D) research strategy, since this would require the development of useful hypotheses before collecting empirical data. Instead, if one wants to learn something about the actor's point of view one has first to enter the empirical field, to establish contact with the people in the field through interviewing or observation and thus collect data. Consequently, qualitative inquiry in most cases starts with observation, recording, listening etc., as also has been mentioned above. This means collecting some-

times large amounts of unstructured textual data, and then hypotheses and theories are developed on the basis of this material.

A qualitative researcher starts his investigation by noticing new and interesting phenomena, and not by searching for evidence for an already developed hypothesis. Of course this must not seduce us into thinking of his mind as a tabula rasa. The researcher always brings some theoretical preconceptions with them. These do not represent hypotheses in the classical sense, that means explicit propositions about empirical facts, but (partly implicit) broad heuristic frameworks which help him to identify and select relevant phenomena, and of course researchers with different perspectives will select different phenomena. Identifying phenomena within one's field notes, protocols or interviews quite often (but not always!) takes the form of "coding".

But this process is quite different from coding in quantitative content analysis, as Charmaz points out:

"Qualitative coding is not the same as quantitative coding. The term itself provides a case in point in which the language may obscure meaning and method. Quantitative coding requires preconceived, logically deduced codes into which the data are placed. Qualitative coding, in contrast, means *creating* categories from interpretation of the data. Rather than relying on preconceived categories and standardized procedures, qualitative coding has its own distinctive structure, logic and purpose." (Charmaz, 1983: 111).

And Becker and Geer explicate the peculiarities of qualitative coding as follows:

"A systematic assessment of all data is necessary before we can present the content of a perspective [...] We have tentatively identified, through sequential analysis during the field work, the major perspectives we want to present and the areas ... to which these perspectives apply. We now go through the summarized incidents, marking each incident with a number or numbers that stand for the various areas to which it appears to be relevant. This is essentially a coding operation, ... its object is to make sure that all relevant data can be brought to bear on a point." (Becker & Geer, 1960: 280-281).

Coding as described by Becker and Geer differs in certain aspects from that in quantitative content analysis. The incidents which are coded do not represent instances or examples of a general phenomenon or fact named by a code; instead the code only refers in a quite vague manner to one of "the various areas to which it appears to be relevant". The purpose is not to condense relevant information with the objective of creating a quantitative data matrix, but to "make sure that all relevant data can be brought to bear on a point".

For these authors the function of codes is clearly restricted to "signposting": codes are stored together with the "address" of a certain text passage, and, drawing on this information the researcher can locate all the possible information provided by the data on the relevant topic. Thus coding in qualitative research is a necessary preparation for the process of systematic comparison. This process forms the basis of the kind of "hypothesis examination" employed in interpretive research, which is quite different from hypothesis testing conducted in an hypothetico-deductive research design. As has been said before a qualitative researcher does not start with readymade and precise hypothesis but develops tentative and sometimes vague hypotheses *ex post facto* from certain parts of the data material, and then further refines and modifies them by drawing on other parts the data material.

This refinement or modification can be supported by code-and-retrieve-software. For example, our hypothetical researcher who developed the hypothesis from a certain interview that there is some sort of hidden competition between different divisions of an organization may be interested in finding all text passages where interviewees talk about the relations between different divisions. Or our other researcher who developed the idea that women and men develop different pain management strategies can try to examine his hypothesis by investigating all text passages where interviewees talk about strategies of pain management. These researchers may find confirming or disconfirming evidence for their hypotheses; in any case they will use the empirical material to further elaborate and modify their tentative conjectures, in line with the concept of Analytic Induction proposed by Lindesmith and Cressey (see also page 42).

For the following reasons, the approach to hypothesis examination adopted by most interpretive researchers is incompatible with the H-D model described above :

Unlike quantitative content analysis, the concepts and categories (the "variables and their values" in the language of quantitative content analysis) which form the basis of a hypothesis are not an integral part of the coding scheme right from the start. The organizational researcher for example did not begin coding the material with the category "competition" right away. A necessary prerequisite for using a certain theoretical concept as the basis of a hypothesis in interpretive research - which incorporates a methodology of discovery - is that this concept has *not* been used to code the raw data. In interpretive, as opposed to hypothetico-deductive, research the researcher must not restrict the scope of the investigation in advance by determining precise categories, since the goal is not to test already formulated hypotheses with empirical material but to generate new ones with the help of empirical material. Coding which support this process means the allocation of text segments to general topics of interest. Text segments that relate to a general topic can then be drawn together to develop hypotheses on the basis of a comparison of these

text segments. Consequently, codes will not be attached to precisely defined incidents in the data but to text segments which are "tentatively classified into the simple content categories we had decided in advance", as, for example, Freidson (1975:271) points out. So Coding and retrieval is nothing more than a mechanical preparation for interpretive analysis which is based on a careful inspection and analysis of raw data (i.e. segments of text) and on their comparison with the purpose of identifying patterns and structures and with the purpose of checking tentative assumptions about these patterns.

In the ongoing process of analysis researchers will usually examine their initial assumptions for the purpose of clarifying, modifying and refining them. But this requires further inspection of the interview transcripts or field protocols themselves, and not of the codes. In many qualitative approaches, especially in those with strong roots in hermeneutic philosophy or phenomenology, such an "interpretive hypothesis examination" would require a thorough fine-grained analysis of textual data. Thereby, further aspects of the phenomenon under study will be discovered through a careful and intensive inspection of the "raw data". In other words, the text itself contains materials that helps to modify the initial hypothesis. Consequently, examining a hypothesis in qualitative analysis is itself an act of interpretation and not an algorithmic procedure based on strict decision rules as with statistical significance tests.

4.3 Two Strategies of Computer-aided Hypothesis Examination

Let us now return to our question about the methodological impact of enhanced retrieval techniques on the research process. As has been said before, some authors claim that facilities for retrieving co-occurring codes can support the process of qualitative hypothesis examination.

It should be clear from the preceding discussion that one has to clearly distinguish between two different strategies, if one uses enhanced retrieval techniques for hypothesis examination, according to the kind of information the computer retrieves when it searches for co-occurring codes: Does the *mere fact of co-occurrence* lead to the rejection or acceptance of a hypothesis, or is it only used to *retrieve the original text segments* which are regarded as the basis for the decision on the hypothesis examined? These two possibilities correspond to the two conceptions of textual analysis mentioned above: one based on a hypothetico-deductive research strategy, the other on interpretive analysis. Both strategies have their individual merits. But since both also have certain mutually exclusive prerequisites, confusing them is likely to be harmful to the research process.

1. In applying a *hypothetico-deductive strategy* the mere fact of a co-occurrence is itself regarded as evidence or counter-evidence for a certain hypothesis. If the researcher proceeds in this way the techniques for searching for co-occurring codes provided by Aquad, NUD•IST or HyperResearch have a similar function to hypothesis testing in statistical analysis. The primary purpose is not to provide the researcher with text segments but to use the information about the co-occurrence of codes in a given document as a basis for decision making. As in statistical significance testing, the decision making process is strictly rule governed and hence algorithmic. So it parallels very much the kind of hypothesis testing which is regularly applied in quantitative content analysis. However, there are certain methodological requirements and limitations to the use of such a hypothesis tester for qualitative hypothesis examination, which are usually taken into account by content analysts. These requirements relate mainly to the nature of codes employed, since the codes must represent Boolean facts if an automatic hypothesis tester is to produce meaningful results. Furthermore, the reliability of the codes used is of utmost importance. But these requirements diametrically oppose the analysis strategy usually applied in qualitative research. In an interpretive analysis strategy codes tend to represent general topics of interest and not precisely defined Boolean facts. Furthermore, hypotheses are not logically stated propositions about the presence, absence or relationship of certain facts, but sometimes vague ideas about the relations between two or more concepts. A hypothetico-deductive strategy where the mere fact of the co-occurrence of certain codes in a given text passage is regarded as evidence or counterevidence can thus rarely be regarded as an adequate strategy of hypothesis examination in interpretive research. There are further reasons why the application of programs like Aquad, HyperResearch, NUD•IST or QCA for hypothetico-deductive hypothesis testing can be seen as a rather dubious endeavour. These strategies have not yet attained the status of statistical hypothesis testing because they do not contain any decision rules, grounded in inferential statistics, for determining when a hypothesis should be rejected or accepted. Instead, they are only suitable for investigating deterministic relationships where the discovery of one contradictory case leads to the rejection of the hypothesis. However, such relationships are extremely rare in social research, let alone in the qualitative field. It would be certainly possible to use methods from inferential statistics to, for example, expand Ragin's proposed process of logical minimization. However, in this case the result would be nothing more than a multivariate model for analysing categorical data. Consequently, one cannot help wondering whether these "new" approaches amount to nothing more than rediscovering the wheel.

There is a further, as yet unsolved, problem with Ragin's concept: that of degrees of freedom. As the number of possible causes increases so does the number of possible

configurations of these causes. With three variables a sample of at least 16 would be necessary to allow each configuration to occur at least once. Therefore, his is a not unproblematic strategy for qualitative researchers who traditionally work with small samples.

Consequently, if a researcher wishes to employ searches for the co-occurrence of codes in a given set of documents in order to test hypotheses within the framework of a hypothetico-deductive strategy, I can only advise them to draw on the tried and tested methods in quantitative content analysis.

2. But the strategies of qualitative hypothesis testing implemented by third generation software as Aquad, NUD•IST, HyperResearch or QCA could nevertheless be extremely useful in qualitative research, if they are used in a quite different way. That would be the case if the results of co-occurring code searches are not regarded as evidence but are used as a heuristic device. The purpose of querying the database would then be to retrieve the original text to which the co-occurring codes are attached. Applying this strategy the result of the retrieval would allow the researcher to determine the meaning of a certain co-occurrence of codes by a thorough analysis of the original text. After the program has retrieved all locations of text passages where segments coded with *cle* co-occur with segments coded with *emo* the researcher would now inspect the original text to answer for example the question: *Has the emotional arousal mentioned by respondent X something to do with the critical life event he describes?* The acceptance of a hypothesis or its rejection (which leads to its further refinement) is not the result of the application of an algorithm (i.e. of a strictly rule-governed process) but is a result of the researcher's interpretation. This corresponds to Lindesmith's and Cressey's method of Analytic Induction, in which the interpretive analysis of interview texts or observations forms the basis for the researcher's decision about a certain hypothesis, while the empirical material, i.e. the textual data, also serves as an information source for generating, refining and modifying hypotheses.

5. Conclusive remarks

This paper started by querying the specific merits and dangers of newly developed computer-aided methods for qualitative research. Let me now try to give a tentative answer. With respect to the *second generation* programs for coding and retrieval, researchers' experiences with using these programs show that they have had a very fruitful methodological impact on the analysis process: by allowing researchers to systematically organize their textual data the software enhances more thorough analyses and reduces the risk that researchers base their results on sparse evidence, that is on a few quotations from some

highly untypical cases. So one has good reason to assume that the continued spread of computer-aided methods for coding and retrieving textual data will enhance the reputation of qualitative research while ensuring the trustworthiness of qualitative findings.

With respect to the newly developed methods of enhanced coding and retrieval, several urgent warnings are necessary: These facilities offer fascinating new possibilities for analysts to "play" with their data and thereby help to open up new perspectives and to stimulate new insights. They can also help to combine qualitative with quantitative methods or an H-D approach with interpretive research strategies. But these possibilities also contain specific dangers because the same technical tool can be used in the context of two totally different analysis strategies: (1.) a strategy in the tradition of interpretive social science whereby textual data are coded for the purpose of hermeneutic analysis of texts and (2.) a strategy in the tradition of content analysis whereby textual data are coded to condense the information contained in them. With the newly developed enhanced retrieval tools for qualitative analysis a qualitative researcher runs the danger of reifying the codes and losing the investigated phenomenon by confusing two analysis strategies. By seeking to test hypotheses without having observed the necessary prerequisites he will easily produce artifacts.

Therefore, I would like to advocate caution when transplanting methodological concepts like "hypothesis testing" or "verification" from one research tradition to another without clarifying their role in the new context.

Notes

1) I am especially grateful to Kate Bird for her comments and critique.

2) It should be mentioned here that by "qualitative research" or "qualitative analysis" I refer to the hermeneutic analysis of textual data in the tradition of interactionist, phenomological or ethnographic approaches and not to the statistical analysis of categorical data which are also often termed "qualitative data".

3) Historical and critical biblical exegesis - which helped in the 18th and 19th centuries to overcome the dominance of dogmatic and literal interpretations of the scriptures - is a good example of this. In particular, the technique of comparing text segments (or "synopsis" as it is called in biblical exegesis) helped with the formulation of a theory about the influences of the four gospels on each other and the order in which they were written, a theory which today is still a generally accepted basis for understanding the New Testament.

Literature

Agar, M. (1991): The Right Brain Strikes Back. In: Fielding, N. G. & Lee, R. M. (eds): *Using Computers in Qualitative Research* (pp. 181-194). Newbury Park: Sage.

Becker, H. & Geer, B. (1960): Participant Observation: The Analysis of Qualitative Field Data. In: Adams, R.N. & Preiss, J.J. (eds): *Human Organization Research: Field Relations and Techniques* (pp. 267-289). Homewood, Illinois: The Dorsey Press.

Charmaz, K. (1983): The Grounded Theory Method: An Explication and Interpretation. In: Emerson, R. M. (ed.): *Contemporary Field Research. A Collection of Writings* (pp. 109-126). Prospect Heights: Waveland Press.

Cressey, D. R. (1950): The Criminal Violation of Financial Trust. *American Sociological Review 15:* 738-743.

Cressey, D. R. (1953/1971): *Other People's Money. A Study in the Social Psychology of Embezzlement*. Belmont: Wadsworth.

Denzin, N. K. (1989): *Interpretive Interactionism*. Applied Social Research Methods Series, Vol. 16. Newbury Park: Sage.

Denzin, N. K. & Lincoln, Y. G. (eds) (1994): *Handbook of Qualitative Research*. Thousand Oaks: Sage.

Dotzler, H. (1995): *Using Software for Interpretive Text Analysis - Results from Interviews with Research Teams*. Paper presented at the Conference SoftStat ´95 (The 8th Conference on the Scientific Use of Statistical Software). Heidelberg, Germany.

Freidson, E. (1975): *Doctoring Together: A Study of Professional Social Control*. Chicago: University Press of Chicago.

Giddens, A. (1976): *New Rules of Sociological Method: A Positive Critique of Interpretive Sociologies*. London: Hutchinson.

Glaser, B. G. (1978*): Theoretical Sensitivity: Advances in the Methodology of Grounded Theory*. Mill Valley, CA: Sociology Press.

Glaser, Barney G. & Strauss, Anselm L. (1967): *The Discovery of Grounded Theory: Strategies for Qualitative Research.* Chicago: Aldine.

Hanson, N. R. (1971): The Idea of a Logic of Discovery. In: Toulmin, S. (ed.): *What I do not believe and other Essays* (pp. 288-300). Dordrecht: Reidel.

Hesse-Biber, S., Depuis, P. (1995): Hypothesis Testing in Computer-aided Qualitative Data Analysis. In: Kelle, U. (ed.): *Computer-aided Qualitative Data Analysis. Theory, Methods and Practice (pp. 129-135)*. London: Sage.

Huber, G. (1995): Qualitative Hypothesis Examination and Theory Building. In: Kelle, U. (ed.): *Computer-aided Qualitative Data Analysis. Theory, Methods and Practice* (pp. 136-151). London: Sage.

Jorgensen, D. L. (1989): *Participant Observation. A Methodology for Human Studies*. Newbury Park: Sage.

Kelle, U. (ed.) (1995): *Computer-aided Qualitative Data Analysis. Theory, Methods and Practice*. London: Sage.

LeCompte, M. D. & Preissle, J. (1993): *Ethnography and Qualitative Design in Educational Research*. San Diego: Academic Press.

Lee R. M. & Fielding, N. G. (1991): Computing for Qualitative Research: Options, Problems and Potential. In: Fielding, N. G. & Lee, R. M. (eds): *Using Computers in Qualitative Research* (pp. 1-13). London: Sage.

Lee, R. M. (1995): *Computer-assisted Qualitative Data Analysis: The User´s Perspective*. Paper presented at the Conference SoftStat ´95 (The 8th Conference on the Scientific Use of Statistical Software). Heidelberg, Germany.

Lindesmith, A. R. (1947/1968): *Addiction and Opiates*. Chicago: Aldine.

Lofland, J. & Lofland, L. H. (1984): *Analyzing Social Settings: A Guide to Qualitative Observation and Analysis*. Belmont, CA: Wadsworth Publishing Company.

Mangabeira, W. (1992): *Contribution to the Closing Session*. Paper presented at the conference "The Qualitative Research Process and Computing". Bremen, Germany.

Mangabeira, W. (1995): Computer Assistance, Qualitative Analysis and Model Building. In: Lee, R. M. (ed.): *Information Technology for the Social Scientist* (pp 129-146). London: UCL Press.

Miles, M. B. & Huberman, A. M. (1984/1994): *Qualitative Data Analysis. An Expanded Sourcebook.* Newbury Park, CA: Sage.

Ragin, C. C. (1987): *The Comparative Method. Moving beyond Qualitative and Quantitative Strategies*. Berkeley: University of California Press.

Ragin, C. C. (1995): Using Qualitative Comparative Analysis to Study Configurations. In: Kelle, U. (ed.): *Computer-aided Qualitative Data Analysis. Theory, Methods and Practice* (pp. 177-189). London: Sage.

Richards, L. & Richards, T. J. (1991): The Transformation of Qualitative Method: Computational Paradigms and Research Processes. In: Fielding, N. G. & Lee, R. M. (eds): *Using Computers in Qualitative Research (*pp. 38-53). London: Sage.

Richards, T. & Richards, L. (1994): Using Computers in Qualitative Research. In: Denzin, N. K. & Lincoln, Y. S. (eds): *Handbook of Qualitative Research* (pp. 445-462). Thousand Oaks: Sage.

Seidel, J. (1991): Method and Madness in the Application of Computer Technology to Qualitative Data Analysis. In: Fielding, N. G. & Lee, R. M. (eds): *Using Computers in Qualitative Research* (pp. 107-116). London: Sage.

Sibert, E. & Shelly, A. (1995): Using Logic Programming for Hypothesis Generation and Refinement. In: Kelle, U. (ed.): *Computer-aided Qualitative Data Analysis. Theory, Methods and Practice* (pp. 113-128). London: Sage.

Strauss, A. L. (1987): *Qualitative Analysis for Social Scientists*. New York: Cambridge University Press.

Strauss, A. L. & Corbin, J. (1990): *Basics of Qualitative Research: Grounded Theory Procedures and Techniques.* Newbury Park, CA: Sage.

Taylor, S. J. & Bodgan, R. (1984): *Introduction to Qualitative Research Methods: The Search for Meanings*. New York: Wiley and Sons.

Tesch, R. (1990): *Qualitative Analysis: Analysis Types and Software Tools.* London: Falmer Press.

Weitzmann, E. A. & Miles, M. B. (1995): *Computer Programs for Qualitative Data Analysis*. Thousand Oaks: Sage.

Wilson, T. P. (1970): Conceptions of Interaction and Forms of Sociological Explanation. *American Sociological Review* 35:697-710.

Znaniecki, F. (1934): *The Method of Sociology*. New York: Rinehart and Company.

Address:

Dr. Udo Kelle, Sonderforschungsbereich 186, Universität Bremen, Wiener Str./FVG-West, D-28359 Bremen, Germany, Tel.: +49-421/218-4168, Fax.: +49-421/218-4153, Email: Ukelle@sfb186.uni-bremen.de

MACHINE-READABLE TEXT CORPORA AND THE LINGUISTIC DESCRIPTION OF LANGUAGES

CHRISTIAN MAIR

To understand the role of machine-readable text corpora in linguistics it is necessary to consider the four possible sources of data for the linguist, viz. (1) the analyst's own introspection/ intuition, (2) more or less systematically conducted elicitation experiments with groups of native speakers of the language studied, (3) collections of authentic spoken or written citations gathered unsystematically, and (4) evidence extracted systematically from a well-defined corpus of texts. After a discussion of the advantages and disadvantages of the various sources of data, I will briefly exemplify recent advances made in the corpus-based description of languages that have become possible as a result of the application of computer technology to linguistics and then go on to present the major databases currently available for the study of English and German.

1. The Linguist's Data

Linguists draw their primary data from four possible sources. These are: (1) the analyst's own introspection/ intuition - fully available only for the mother tongue, (2) elicitation experiments - ranging from informal polls among friends and colleagues to systematic test batteries used with representative samples of speakers, (3) authentic spoken or written citations collected randomly, and (4) well-defined text corpora from which the relevant evidence can be extracted exhaustively and systematically. The source of data chosen limits both the type of question that can be asked and the results likely to be obtained.

Consider, for example, the English "medio-passives" illustrated by (1):

(1) This book reads well also in translation.

Officials bribe easily in some countries.

* The figures will understand better if presented in a table.

For the benefit of an audience consisting chiefly of non-linguists, the "medio-passive" can be defined as a grammatical structure that is active in form (i.e. the verb *reads* has the same form as in *she reads lots of books*), but passive in meaning (i.e. the first example can be rephrased as *this book can be read well also in translation*). English mediopassives have aroused linguists' interest because they are a recent and spreading innovation and also because English affords a greater variety of such constructions than other European languages.

Of the three sentences given in (1), a native speaker consulting his linguistic intuition will very likely accept the first as normal, the second as possible but somewhat strange but reject the third (hence, following standard linguistic conventions, the asterisk in front of this example). If asked, the native-speaking informant could come up with a few dozen more examples of the construction, which is probably enough for a first approximative linguistic description of the English medio-passive.

However, here as in many other areas of the grammar, there is a grey zone in which native speakers' judgments are unreliable. Consider:

(2) Children educate less easily nowadays than they used to.

Most native-speaking informants will hesitate to give an immediate judgment on this example and resort to *ad-hoc* explanations of the kind: "well, I could imagine a journalist using it," "maybe it's American" (if they themselves are British) or "sounds British" (if they are American).

Obviously, it is precisely this type of phenomenon that lends itself easily to investigation on the basis of elicitation and corpora. Thus, a hundred British or American speakers could be asked for their opinions on:

(3) *East Enders* will screen weekly at 8 p.m. from next Monday.

If the results remain unclear, the number of informants could be increased, or they could be re-grouped according to level of education, age or other potentially significant variables. One problem, however, will remain: what is tested in such elicitation procedures is informants' metalinguistic judgment, which does not necessarily reflect their spontaneous language use. Tests designed to elicit performance are very difficult to construct - how, for example, could an informant be made to actually say or write the structure illus-

trated in (3)? And even in cases where this is possible, it is far from clear whether such elicited performance is the same as spontaneous language use in a natural communicative context.

In this particular case, it is thus the corpus-linguistic approach that will bring us closest to an answer. Analysing machine-readable newspaper corpora from the U.K., the US, Australia and New Zealand, Marianne Hundt (personal communication) finds that the medio-passive use of *screen* originated in the latter variety and is still largely specific to it. Note, however, that one thing which is still unclear is whether this usage is typical of New Zealand English usage in general or restricted to the professional jargon of a particular sub-group of speakers such as, for example, media professionals. At the end of our illustrative analysis, we are therefore thrown back to square one and would probably have to continue by asking native speakers of New Zealand English for their intuition on the matter.

If this little introductory example suggests that in linguistics as well as in many other fields the empirically most adequate descriptions result from methodological pluralism and eclecticism, this is a message which the present writer can endorse fully. In what follows, however, I shall confine myself to an exclusive discussion of the advantages and limitations of the corpus-based approach to the study of languages.

2. The Corpus-Linguistic "Take" on Language

Corpus-based linguistics had a first heyday in the late nineteenth and early twentieth centuries. The renowned Danish Anglicist Otto Jespersen, for example, counted - or, to paraphrase his own words, asked one of his pupils to count - continuous forms of verbs (*I am reading*, etc.) in extracts from two successive translations of the Bible in order to show how such forms had spread in the language from the Middle English period to the time of his writing (1909-49: IV, 177).[1] Of course, the "Shakespeare-corpus" and the "Chaucer-corpus" have long been concordanced and used for literary as well as linguistic research. Charles C. Fries, a prominent representative of the American structuralist school, wrote a grammar of English based on a corpus of letters and telephone conversations. Today, we are awed both by some of the results of such work and by the drudgery that must have gone into it. The inevitable tedium associated with work on them probably explains why corpora went out of fashion for a time.

It was the advent of computational storage and retrieval methods that gave corpus linguistics a new lease of life from the nineteen-sixties onwards. After relatively timid beginnings, we are now seeing a veritable boom in the field, which is due to rapid developments in hardware and retrieval software, the availability of more, larger, and more diversified corpora, and also to the fact that more and more corpus-linguists are beginning to realise that the statistics and examples they dig up are interesting to the linguistic community only to the extent that they shed light on important empirical and theoretical issues.

Today, corpus-linguists have gathered a formidable body of evidence that may eventually force a reorientation in linguistic theory. While in Chomskyan linguistics and related schools of thought which have dominated theoretical debates in the field for the past thirty years, grammatical structure is modelled as an autonomous formal algorithm, corpus linguistics emphasises the fuzziness of all grammatical categories, and the interdependence between structure, meaning and context. The distinction between underlying linguistic "competence" (Chomsky) - a neat and tidy system accessible to introspection - and "performance" - its imperfect realisation in communication - has become increasingly difficult to uphold. The analysis of large masses of data as they are available in machine-readable corpora reveals that there is patterning and order also on the level of performance, and that this needs to be taken into account if we want to understand the nature of language.

Where once there was poverty of data - the changes used to be rung on standard examples of the type *John sees Mary* - there is now an abundance of data. Consider, for example, the following authentic attestations of the verb *see* which were all culled from the 1995 electronic edition of the *Guardian* (1 - 31 March) with minimal effort. They are outside the scope of what general reference grammars and dictionaries have to say about this verb and thus pose an immediate challenge for linguistic analysis:

(4) Although the government is committed to seeing justice done, there are few qualified people to work in the judiciary to deal with the prisoners. (11 March 1995)

(5) Already there are signs that the electrical and machine tools sectors, two of the pillars of the German economy, are seeing their order books affected by the exchange rate. (17 March 1995)

(6) "I am very interested in seeing cable companies provide local channels for community access." (30 March 1995)

Here the verb *see* is not used in its established literal or metaphorical meanings but as a grammatical device. A government that wants to "see" justice done does not really want to see anything at all; it merely wants justice to be done. If someone sees their order books affected by a change in the exchange rate, the emphasis is not on the act of physical or mental perception as such: this is just another way of saying that their order books are affected. Likewise, "to be interested in seeing cable companies do something" is the same as saying "to be interested in their doing something." The use of the verb *see* merely results in a grammatical restructuring of the sentences in question that highlights the position of the grammatical subject as an affected entity, because the gratuitous use of the verb introduces a note of physical substance and human activity into an otherwise abstract state of affairs.

The process of "grammaticalisation" - to give it the name commonly applied to such processes in linguistics - has proceeded even further in the following two examples of nonliteral uses of the verb *see*:

(7) His almost obsessional devotion to horseracing is the more remarkable seeing that his beloved father had striven to keep him away from racing in any shape or form. (12 March 1995)

(8) Not even the Kennel Club, which had no comment to make. Perhaps that's just as well seeing what a hash it makes of things when it does deign to comment. (19 March 1995)

This is not the verbal participle *seeing* that we have in a sentence like *seeing a huge crowd approach, he withdrew.* Rather, *seeing* has become a conjunction, a grammatical operator introducing relevant background information and paraphrasable as "in view of the fact."

This is not a unique development. In the course of its history, English has seen many verbs turn into grammatical operators, for example *concerning*, *regarding*, or *if* (from the Old English imperative form for *giefan*, "give"). The difference between these examples and *seeing* is that the split between the lexical verb and the grammatical operator is not yet complete, and the meaning(s) of the emerging grammatical operator(s) are therefore difficult to pin down precisely. Grammaticalisation is a gradual process, in which the statistical proportions of lexical vs. grammatical uses of a word slowly shift from generation to generation. Data from individual native speakers' introspection are largely irrelevant to a description of the phenomenon, and it is therefore very likely that with the availability of more and more machine-readable corpora covering more and more lan-

guages and registers, the study of grammaticalisation processes will be raised to a qualitatively new level in the near future.

I have mentioned an example from the field of grammaticalisation because it makes a point already made above. The analysis of corpora yields the best results when the new data and technology are used to address existing theoretical issues, and grammaticalisation is just one such case. The fact that lexical items are worn down to grammatical operators in the course of language history is a linguistic universal and has never escaped serious linguists' notice. In fact, the term "grammaticalisation" was coined and defined by Meillet as early as 1912. What was lacking were the rich and easily accessible data bases that would have been required to refine and flesh out a promising hypothesis with the necessary empirical detail. What in the study of language change in Middle English will always remain a dream - matching corpora documenting the state of development in the language every 20 or 30 years for reliable quantitative and qualitative analysis - has now become a reality, and genuine advances in linguistic scholarship are possible.

To put it as briefly as possible, machine-readable corpora are a superior source of data for the linguist for four reasons. The first two are practical in nature, the remaining two concern a more central issue, namely the goals of linguistic research:

(1) Machine-readable corpora make it possible to retrieve large amounts of linguistic data with minimal effort, which allows the exploration of promising as well as not-so-promising working hypotheses in a reasonable span of time. A year's work may be wasted if a large corpus is scanned manually for data which later turn out to be useless; a few hours' worth of work are lost if a similar search is done by computer.

(2) Access to most corpora is not restricted to a few individuals but open to larger sections of the research community. Linguists working on the same material can thus build on each other's results, which leads to co-operation and cumulative progress of a kind not typical of the field as a whole.

(3) Machine-readable corpora are superior sources of data because they present data in their original textual and situational context, sharpening our awareness of the flexibility with which grammatical rules and categories are applied in practice and of the many interdependences between form and meaning or language and context. Close scrutiny of individual examples in context constitutes the qualitative aspect of corpus-based linguistics.

(4) Machine-readable corpora are superior sources of data because they make it possible to analyse the data statistically where, as in the case of grammaticalisation or in the study

of regional, social or text-type specific variation, such analysis is desirable. This it the quantitative aspect of corpus-based work.

In order to give interested outsiders a flavour of what corpora are typically used for these days in linguistics, I would like to conclude this section by briefly commenting on the seventeen contributions published in a recent volume of conference proceedings (Fries, Tottie, Schneider, eds. 1994).

Three contributors report on corpora they are themselves compiling. The first is ARCHER, a somewhat laboured acronym based on the project-title "A Representative Corpus of Historical English Registers." The texts in this corpus are divided into historical compartments of 50 years each, with the focus being on British English but three periods (1750-1799, 1850-1899, 1950-1990) also providing American material in order to make possible the systematic study of regional as well as historical variation. All told, the corpus contains about 1.7 million words and, by present standards, has thus to be included among the small specialised corpora in the field. As an illustration of the type of problem ARCHER could be used to investigate, the compilers show that there is a drastic increase in information-orientation in medical writing after the middle of the nineteenth century, which can be taken as a symptom of the discipline's redefining itself as a hard science. The second paper devoted to corpus building is a report on the progress of the Hong Kong component of the International Corpus of English (on which see below), and Johansson/ Hofland introduce their projected English-Norwegian parallel corpus, which is supposed to benefit not only theoretical linguists but also bilingual lexicographers, language teachers and translators.

It is not without apprehension that I move on to a selective survey of the fourteen papers following, because the type of problem studied by linguists often meets with baffled incomprehension outside the field. But here is the list of what we waste our time doing, for what it is worth.

As I have pointed out above, one strong point of corpora is that they show how artificial the dividing line between grammar and the lexicon really is and how much of language consists of habitually used recurrent word combinations. Henk Barkema pursues this line of inquiry and wants to find out which idioms are inflexible multi-word lexical items and which allow limited modification. *Cold war*, for example, turns out to be of the latter type, with attested variants including *not-so-hot wars*, *melting cold wars*, *periods of hot and cold civil wars*, and so on. Eeg-Olofsson and Altenberg study discontinuous recurrent word combinations (frames like *in_of* or *in_with*) in a corpus of spoken English and - predictably - show that the most frequent way of filling these slots is to produce more prefabricated building blocks (*in terms of, in touch with*) rather than creatively coined

novel expressions. Peters compares variant past tense and participle forms for verbs (e.g. *dreamed/ dreamt*) to determine whether Australian norms are closer to British or to American ones in this part of the grammar. A whole cluster of papers is given to attempts at statistical identification of language change (Nevalainen on adverb derivation, Raumolin-Brunberg on the placement of adjectival modifiers in Late Middle English), text-types or stylistic registers (for example Svartvik/ Ekedahl/ Mosey on "public speaking"). Somewhat surprisingly, the compilers and the users vastly outnumber the computer-science contingent in this volume, as questions of tagging and parsing, that is the automatic grammatical analysis of natural language data, are touched on in only one paper (Voutilainen/ Haikkilä).

3. The Major Resources for the Corpus-Linguist Working on the English Language

In what follows I will discuss the most important English-language corpora, focussing chiefly on those that can be installed on personal-computers and are thus available for desk-top research. For a fuller picture, the reader is referred to Taylor/ Leech/ Fligelstone 1991, who list the resources more completely, and Altenberg 1991, a bibliography which documents the major research done using them. Annual updates are provided in the *ICAME Journal* (Bergen, Norway).[2]

The corpora to be discussed fall into two groups: (a) those that are in the range of roughly one million words, carefully sampled and proofread, inevitably aging, and - owing to their limited size - chiefly of use for the study of the most frequent words and grammatical structures in the language, and (b) large and sometimes open-ended collections of text in which proofreading and principled sampling techniques have to some extent been sacrificed on the altar of size.

The prototype of the type (a) corpus is the Standard Sample of Edited American English, named the Brown Corpus after the institution the project was based at. It contains 500 samples of about 2,000 words each, spanning 15 textual genres from press reportage to various types of lowbrow fiction. The British LOB (for Lancaster-Oslo/Bergen) corpus followed, with the new opportunity for systematic research into British-American differences soon yielding an impressive body of research literature. Matching corpora documenting second-language Indian English (Kolhapur), Australian English (Macquarie) and New Zealand English (Wellington) followed suit. The problem was that while the first two corpora, namely LOB and Brown, contained only texts first published in 1961, the later clones included material from the nineteen-seventies and eighties, thus introduc-

ing a most unwelcome distorting factor in the shape of possible linguistic change over this period of time. In order to remedy this difficulty, the present writer decided to compile two new British and American corpora which are to match the originals as closely as possible in size and composition except that the texts included were published in 1991 and 1992. The projects await completion in 1996 and are known in the linguistic community by their somewhat facetious working titles FLOB and Frown (for Freiburg-LOB and Freiburg-Brown). On completion of the new corpora, it will be possible to study systematically not only regional variation between written British and American English but also linguistic change in progress. A question which it will be possible to ask, for example, will be to what extent the grammar of British English has been influenced by American usage over the past thirty years.

The Survey of English Usage corpus (initiated in the pre-computer era by Sir Randolph Quirk, University College London) is a one-million word corpus which contains a sizable amount of surreptitiously recorded spontaneous speech, part of which was later made available in machine-readable form and published in prosodic transcription (Svartvik/ Quirk, eds. 1980). This corpus of English conversation has not been surpassed as a databasa for research on spoken English so far.

The latest venture in the small-corpus field is the International Corpus of English (Sidney Greenbaum, London), which aims to document spoken and written English of the nineties from all major native-speaking and second-language communities. The British component of ICE will be available to the research community shortly.

A pioneering effort in the development of vast and open corpora was the COBUILD corpus (John Sinclair, Birmingham), which from the beginning was planned as a joint venture between academia and the publishing business and has so far resulted in a number of dictionaries and other reference and teaching materials. It has sprouted a number of successor projects, which - like the original - are accessible to the general public with some difficulty only. Most of these ventures are biased towards written language, and it might be argued that they have become the victim of extremely rapid progress as meanwhile vast quantities of machine-readable English, continue to pile up every year "by themselves", as it were, because practically all the major British and American newspapers and journals offer machine-readable editions on CD-ROM.

The crowning achievement in this tradition is undoubtedly the recently published British National corpus, which contains over 100,000,000 words of British English from spoken and written texts and in which - and this is a trailblazing innovation - every word has been automatically tagged for part-of-speech membership. It is the product of a consortium bringing together major dictionary publishers, academic institutions (chiefly the

University of Lancaster) and the British Library (Research and Development Division). It probably offers the best of both the "small" and the "large" corpus worlds, because it is distributed at a very low price and yet offers a hitherto unattained amount of clean and orderly data. Also, it goes some way towards redressing the major drawback in most previous projects, namely the under-representation of spontaneous speech.[3]

It is impossible for me to give a similarly detailed survey of corpora available for the study of other languages. However, one resource which needs to be mentioned is the Multilingual Corpus published in 1994 as a compact-disc by the European Corpus Initiative of the Association for Computational Linguistics. People interested in computer-based work on German language and literature will find annual updates in the first yearly issues of the journal *Germanistik*. The major centre of corpus-based descriptive work on modern German is, of course, the Institut für Deutsche Sprache (IDS) in Mannheim.

4. Conclusion: Corpus - Linguistics and Neighbouring Fields

The greater part of corpus-based research in linguistics concerns questions and problems that are specific to the discipline and of little interest to outsiders. However, the corpora I have described are available to researchers from other fields for their own purposes. In some areas a dialogue between corpus-linguists and other scholars using the same resources is bound to yield interesting results. In conclusion, I should like to mention three areas in which such interdisciplinary dialogue has in my opinion been long overdue.

In philosophy, Ludwig Wittgenstein started a tradition in which philosophical inquiry was to be preceded by a close scrutiny of the natural-language uses which crucial terms of the analysis were put to. Philosophers of the natural-language school have usually relied on their introspection to establish unreflective natural uses of words and expressions. Without being polemical, however, one might ask whether a trained philosopher's intuition is the closest we can get to current communicative practices in a community of speakers. An analysis of thousands of actual uses of a critical expression as could be culled from suitable corpora is certainly a better indicator of community consensus in usage.

Lexicographers working on large and continually updated corpora are in a perfect position to record the emergence, spread and establishment of new words. Developments in the vocabulary of a language, however - be they the introduction of new words or subtle changes in the meanings of existing ones - frequently mirror changes in society and speakers' attitudes. Take, for example, the increasingly frequent use of the adjective *aggressive* as a positive evaluation of behaviour (as in "aggressive negotiaton tactics", which presumably means "successful tactics," or "an aggressive and ambitious business student"). To the linguist, this is one more example of semantic change of the ameliorative type, in which the meaning of a word loses some of the negative connotations originally associated with it; to the social scientist it may be an indicator to value changes in the community.

To end with a self-evident example, one might point out that the quality of a computer-scientist's or an artificial intelligence researcher's work on natural-language processing is in direct proportion to the corpora that he or she can use as a testing ground for tagging and parsing programs.

Notes

1) Jespersen is here quoted as an illustrative example of corpus-use before the advent of computers. His grammar is otherwise largely based on the author's inexhaustible collection of citations, which he amassed over a lifetime of diligent scholarship, and can thus serve as a perfect example of the third of the four data-gathering strategies mentioned above.

2) ICAME, P.O. Box 53, N-5027, Bergen, Norway, the Association for Computational Linguistics, c/o D. E. Walker, Bellcore, MRE 2A 3-79, 445 South Street Box 1910, Morriston, NJ 07960, USA, and Oxford University Computing Services, 13 Banbury Rd., Oxford OX 2 6 NN, England, are the three major clearing-houses for up-to-date information on resources available and other logistical matters in the field of English corpus-linguistics.

3) To mention one of the more ingenious ways of doing so, for example, certain demographically representative individuals were wired with recording equipment during a set period of time in order to document and obtain a sample of their most authentic speech.

Literature

Aijmer, K. & Altenberg, B. (eds.) (1991): *English corpus linguistics*. London: Longman.

Altenberg, B. (1991): A bibliography of publications relating to English computer corpora. In: Johansson, St. & Stenström, A.-B. (eds.): *English computer corpora: Selected papers and research guide* (pp. 355-396). Berlin: Mouton de Gruyter.

Atkins, B.T.S., Levin, B. & Zampolli, A. (1994): Computational approaches to the lexicon: An overview. In Atkins, B.T.S. & Zampolli, A. (eds.): *Computational approaches to the lexicon* (pp. 17-45). Oxford: OUP.

Butler, C. (ed.) (1992): *Computers and written texts*. Oxford: Blackwell.

Fries, Ch. (1940): *American English grammar*. New York.

Fries, U., Tottie, G. & P. Schneider (eds.) (1994): *Creating and using English language corpora*. Amsterdam: Rodopi.

Jespersen, O. (1909-49): *An English grammar on historical principles*. 5 vols. Copenhagen: Munksgaard.

Sinclair, J. (1991): *Corpus, concordance, collocation*. Oxford: OUP.

Smith, G.W. (1991): *Computers and human language*. Oxford: OUP.

Svartvik, J. (ed.) (1992): *Directions in corpus linguistics*. Berlin: Mouton de Gruyter.

Svartvik, J. & Quirk, R. (eds.) (1989): *A corpus of English conversation*. Lund: Lund University Press.

Taylor, L., Leech, G. & Fligelstone, St. (1991): A survey of machine-readable corpora. In: Johansson, St. & Stenström, A.-B. (eds.): *English computer corpora: Selected papers and research guide* (pp. 319-354). Berlin: Mouton de Gruyter.

Address:

Professor Dr. Christian Mair, Albert-Ludwigs-Universität Freiburg, Englisches Seminar I, Institut für Englische Sprache und Literatur, Kollegiengebäude IV, D-78095 Freiburg, Germany, Tel: +49-761/203-3336, Fax: +49-761/203-3340

PRINCIPLES OF CONTENT ANALYSIS FOR INFORMATION RETRIEVAL SYSTEMS: AN OVERVIEW

JÜRGEN KRAUSE

Unquestionably, the content analysis which has emerged as part of Information Retrieval Systems (IRS, e.g. literature databases) over the past 20 years has much in common with the content analysis used by linguists or in the social sciences. However, its intrinsic value stems from the special context in which it is used:

a) Close interdependencies link the selected content analysis with the retrieval situation. The user's retrieval strategies, which are intended to obtain information relevant to the current problem situation, and the available aids (e.g. expansion lists or user-friendly browsing tools) affect the efficacy of some analysis techniques (e.g. noun phrase analysis from computer linguistics) to a considerable extent.

b) Normally, a commercial IRS handles mass data, thus necessitating the use of a reduced content analysis even today. Full morphological, syntactic, semantic and pragmatic text analyses are unthinkable simply for efficiency reasons but also for knowledge reasons. Content analysis in IRS is therefore a component part of a special type of restricted system which obeys its own laws.

Against the backdrop of these considerations, forms of content analysis in present-day commercial retrieval systems are studied and promising expansions and alternatives are proposed.

1. Introduction

The objective is to show possible approaches for improving retrieval functions of IRS based on the state of the art now attained in commercial systems and practice-oriented

developments in research, and to determine the advantages and disadvantages of individual solutions. Since content analysis measures can be frequently exchanged with those on the retrieval side, both the aspect of content analysis and retrieval will be examined. A certain type of content analysis may therefore only be chosen to organize the retrieval algorithm efficiently. A simple example is the retrieval function of truncation. It is largely superfluous if compound splitting and basic form reduction are used in content analysis. However, compound splitting and basic form reduction can also be replaced by equivalent generation methods during research. In an ideal situation, the user does not notice whether an algorithm expands the user's search word to include all word forms or whether the word forms of the document are reduced to basic forms when descriptors are allocated.

Commercial text IRS are now primarily based on intellectually or automatically determined descriptors which are researched with or without additional thesaurus relations by means of Boolean algebra. The following comments are restricted to this type of research and the overcoming of its inherent weaknesses by computer-linguistic and quantitative-statistical methods. In addition, information science is aware of other basic types of retrieval which each attract other or modified content analysis methods. The best-known and most widespread additional types of search are:

a) Search path organized on a hierarchical basis

Due to an ever greater restriction on selected generic terms, the user is guided to his target data by means of a path specified by the system. All selection alternatives are provided on the user interface. Active noting features are not required. The user need not search for his own terms, instead he (she) makes a selection from a displayed repertoire.

However, these advantages are accompanied by marked disadvantages: the low quantitative limits of this solution must be taken seriously. If the hierarchical structure (especially depth) increases, clarity on screen is quickly lost. It depends a great deal on the (normally objective) conditions of the area of application whether a hierarchical system can be considered. Hierarchical access alone is only adequate very seldom.

In the case of hierarchical systems, the advantage of content analysis primarily lies in the intellectual construction of the flow chart and allocation of the individual documents to the nodes.

b) Hypertext relations

Kuhlen (1991) extensively examined possibilities and limits of hypertext systems and their realization forms. Hypertext realization should be regarded as an additive element

of conventional descriptor systems in the same way as hierarchical access - supplementary to special search situations - can improve descriptor systems.

In addition to a) and b), the information science discussion is aware of a number of other types of research. In Bates (1989:412), for example, the traditional descriptor system is only one of seven basic types which would all have to be supported in an ideal text research system:

> "1. *Footnote chasing* (or 'backward chaining' ...). This technique involves following up footnotes found in books and articles of interest, and therefore moving backward in successive leaps through reference lists...
>
> 2. *Citation searching* (or 'forward chaining'...). One begins with a citation, finds out who cites it by looking it up in a citation index, and thus leaps forward.
>
> 3. *Journal run*. Once, by whatever means, one identifies a central journal in an area, one then locates the run of volumes of the journal and searches straight through relevant volume years.
>
> 4. *Area scanning.* Browsing the materials that are physically collated with materials located earlier in a search is a widely used and effective technique. Studies dating all the way back to the 1940s confirm the popularity of the technique in catalog use.
>
> 5. *Subject searchers in bibliographies and abstracting and indexing (A & I) service*. Many bibliographies and most A & I services are arranged by subject. Both classified arrangements and subject indexes are popular. These forms of subject description (classifications and indexing languages) constitute the most common forms of 'document representation' that are familiar from the classic model of information retrieval discussed earlier.
>
> 6. *Author searching*. We customarily think of searching by authors as an approach that contrasts with searching by subject. In the literature of catalog use research, 'know-item' searches are frequently contrasted with 'subject' searches, for example. But author searching can be an effective part of subject searching as well, when a searcher uses an author name to see if the author has done any other work on the same topic."

The expansion of the possible types of research in an IRS is an important subject for the further development of current commercial approaches. Due to reasons of space, this subject must unfortunately be omitted along with components of intelligent Information Retrieval (IR) which were already discussed in Krause (1992).

2. Principles of content analysis and information retrieval

By way of introduction, it is necessary to clarify some fundamental principles and problems of IR which are all ultimately rooted in an indistinct way at all system levels (Krause, 1990). This generally occurs - although content analysis is a main subject - from the viewpoint of search, as every problem appears to the user as a search and interaction problem. He (she) also experiences all content analysis measures as indirect impacts on his (her) search formulation, which is why conceptualization of the entire process should be seen from this aspect.

2.1 Descriptor search using Boolean algebra for document retrieval

Boolean algebra can be used to establish a link between queried terms by means of the logical operators AND, OR, NOT. These are often supplemented by formal additional techniques such as truncation (right, left, inward truncation) or neighborhood search (for closer definition of the AND operator) which function exclusively through exact-pattern-match processes.

During document retrieval with Boolean algebra, the user can normally fall back upon two types of descriptor regarding a document:

a) Aspect-related descriptors such as name (obligatory), organizational code, author and (modification) date and

b) Free descriptors (= keywords) which can also be obtained from the already existing stock of descriptors by marking the list entries (= connection with the model character).

From the viewpoint of content analysis (indexing), the descriptors - irrespective of any syntactic or hierarchical reference (more generally: without any relationing with each other with the exception of some aspect statements) - will characterize the contents of the document. Users utilize the same unrelationed, content-identifying terms in the search, thus avoiding the mentioned difficulties with hierarchical access.

2.1.1 Thesauri

Commercial descriptor systems normally permit a number of general connections between individual terms (synonyms, associations, generic terms, subterms, etc.). Thanks to these global relations, a list of descriptors turns into a thesaurus. If the thesaurus contains generic terms and subterms, an attempt is made in a descriptor system containing no

second search path to integrate the advantages of an hierarchical system in the list of descriptors (cf. DIN 1463).

Nowadays, there is no longer any doubt concerning the general effectiveness of thesauri which are generated intellectually or semi-automatically to improve the widest range of IRS (see, for example, the Darmstadt indexing approach: Lustig, 1986; Fuhr, 1988; Biebricher et al., 1986).

2.1.2 Intellectual versus automatic descriptor determination

The discussion concerning both these basic types is as old as the first IRS used in practice. For example, the Informationszentrum Sozialwissenschaften (IZ Bonn, 1994), develops its literature databases fully intellectually on the basis of a thesaurus which is organized partially according to a hierarchy. The indexer takes all allocated terms (exception: additional field for "free" terms) from a thesaurus which must be constantly updated. The terms contained in the thesaurus form an integrated semantic system. The advantage of these terms is that the indexing depth can be controlled during generation of the thesaurus and semantic standardization is enforced at the level of indexing. However, the terms in the restricted, specified vocabulary can "lose" their colloquial semantics and can be almost interpreted in formal language. This characteristic of controlled, intellectual indexing becomes most apparent when an indexer does not find a desired term in the thesaurus.

Purely intellectual indexing based on controlled thesauri produces excellent results in some areas. However, this approach is hardly used any more in practice for large data stocks since the costs of intellectual processing are regarded as too high. Users of these systems also do not normally consult the list of keywords, but formulate their query directly. If the thesaurus is not constantly updated to include new (fashionable) terms in a specialist area, this results in excessive discrepancies between the terms selected by the users and the thesaurus structures.

Moreover, it has not yet been possible to the best of my knowledge to prove the postulated advantage of intellectual analysis using a controlled thesaurus by means of corresponding evaluation.

Intellectual processing has largely been replaced in commercial systems by automatic free-text methods in which thesauri and, if necessary, computer-linguistic methods (e.g. basic form tracing) are only still used in sub-areas to resolve the linguistic diversity of the starting texts.

The DATEV databases are, for example, an extreme example of the "pure form" of traditional, automatic free-text systems (cf. DATEV, 1994). Except for some aspect-related descriptors, only a stop word list regulates the selection of descriptors. The JURIS databases (GOLEM/PASSAT base, cf. Möller, 1993) are another example. It is interesting to note here that after more than fifteen years practical experience, intellectual analysis, which was vehemently rejected at the start of development, is again regarded as a cure for the empirically observed, poor retrieval performances of JURIS (cf. Wolf, 1992; Möller, 1993).

2.1.3 Summary

With all of the above-mentioned methods, users often gather during their work extensive heuristic knowledge on what descriptors can be formulated adequately or as search queries. Users also become specialists regarding Boolean retrieval. The basic principle of the inverted list also produces good response times because the actual data stock need only be accessed when the user wants to examine document texts.

However, these advantages of conventional Boolean retrieval are contrasted by marked disadvantages.

Before they are discussed in detail, the theoretical basis of the so-called standard model will be explained more fully.

2.2 Standard model of traditional information retrieval

This is also called the Salton paradigm (cf. Belkin, 1993:57). Basically speaking, it remained the starting point of any suggested improvement regarding IR up to the 1980s. The central element in this model is the exact pattern match method which combines the terms selected by the user with those in content analysis. As shown above, only the logic operators AND, OR, NOT, brackets and formal additional techniques such as truncation or neighborhood search are available (e.g. restricting AND to the sentence through the corresponding context operator).

The methods described in the next sections improve the pattern match method by eliminating some obvious weak points (in particular, section 3) or replacing Boolean algebra by quantitative-statistical procedures (cf. section 4), but do not affect the general validity of the basic model.

Its main weaknesses become most apparent when the user modifies the query after the system has supplied the initial results:

> "Concern with representation of the information need has typically arisen after the process of judgement, which is typically to be performed by the user, as an estimate of the potential relevance of the text to the information need. The results of the judgement process are then used by the system to modify the query, or, occasionally to modify the text representations. This process of... 'relevance feedback', is perceived ... as an attempt to gain the 'best' possible representation of the user's query ... that is, to improve the representation so that the comparison process will work most effectively. It is important to recognize that, in this model of IR, the person involved in IR is seen as a user of the system, standing outside of it. Involvement of the user with the IR system is minimal and interaction (in the form of the judgement process) is seen as ancillary to, and only in support of, the representation and comparison processes." (Belkin, 1993:56).
>
> "The force of these assumptions is ... to devalue or even ignore the significance of interaction of the 'user' with the texts; and to provide support for only one form of information seeking behaviour, that associated with searching for some well-specified item. Additionally, through the privileged position of comparison and representation, and the assignment of responsibility for these activities to the system, the standard view of IR leads to strong control by the system of the entire IR process, and the consequent lack of power or control by the user" (Belkin, 1993:57).

In the discussion of the principles examined here, this global criticism is used more as a general guideline to remedy the weaknesses of the standard mode by additional measures. Only the discussion - mainly omitted for reasons of space - concerning components of an IR and that concerning additional search types force us finally to leave the model (cf. section 1).

3. Computer-linguistic methods as a supplement to free-text retrieval

Improperly from the viewpoint of computer linguistics, traditional free-text retrieval reduces content analysis to a symbol-oriented analysis (cf. Knorz, 1994). Documents are regarded as the stringing together of chains of symbols which are separated by blanks or punctuation marks. Every morphological, syntactic or semantic item of information contained in the text is classified as negligible, which is the reason why it is not analyzed.

The designers of these systems (hereinafter described by means of examples in German) naturally know that the omitted morphology, syntax and semantics shorten content

analysis, i.e. for the most part irreversibly: syntactic information in the documents is no longer available - once erased during document recording - when the query is made (cf. the classic example: *Suche nach griechischen Schiffen, die römische Häfen anlaufen* (look for Greek ships which call at Roman ports*)* versus *römische Schiffe, die griechische Häfen besuchen* (Roman ships which visit Greek ports). The methods developed for this purpose are not dependent on language and can therefore be used internationally. Due to their simplicity, they can be developed in an extremely robust manner for mass data. They are also fully automatic, quick and cheap. Symbol-oriented, general methods on the query side will counter the lack of analysis depth at least in the area of morphology and, at times, in syntax. Truncation and context operators are actually operations without "linguistics", but the developers of the IRS are confident that the user has this knowledge (e.g. *Haus** for all compound words such as *Haustür, Hausverwaltung,* etc.) and uses it adequately in search strategies. By resorting to user knowledge, retrieval-side truncation will therefore solve the problem of word form combination, compound splitting and derivation combination, and intelligent use of context operators will replace the lost references of sentence structure.

3.1 Methods available

The computer-linguistic argument assumes that the reduction of content analysis to a purely symbol-oriented, non-language-dependent dimension is responsible for the majority of research problems. The language-dependent rule systems researched by linguists are therefore integrated. They replace the user's linguistic skills which he (she) must transfer during traditional free-text research to the symbol-oriented help operations (truncation, context operators) which are by no means suitable for this purpose. The quality of the exact pattern match method of traditional IR systems will be improved by:

a) returning the word forms in the text to their basic forms or expanding the word forms of the user query to all related basic forms. The recognition of abbreviations can be classified here as a special case.

e.g. *Stall: Stalls, Ställe, Ställen - Thema: Themen, Themas, Themata*

b) Splitting compound words into their constituent parts

e.g. *Druckerzeugnis: Druck-Erzeugnis* versus *Drucker-Zeugnis*

c) Combination of derived terms

e.g. ADJ *lieblos* / NOUN *Lieblosigkeit* (*-keit* turns adjectives into nouns);

Formatierung / Format / formatieren

Rang er nach Luft: basic form *Rang* or *ringen*

d) If the structure is given related terms (multi-word terms, complex descriptors, verb prefixes such as *gehört ... zusammen*)

e.g. *fing ... an ...*: basic form *fangen* or *anfangen* - *hielt ... den Atem an*

• natürliche und juristische Personen - kalter Kaffee (= Spezi) versus kalter und abgestandener Kaffee

e) Hyphenated part-word deletions e.g. *Haus- und Hofwirtschaft*

f) Phonological-graphematic translation of proper names and spelling checker

Spelling checkers are a standard feature of text-processing programs and are supplied by special companies for a large number of languages. Curiously, most information systems do not have a spelling correction feature for the search term. It should be as natural as a check on the entered texts. This is especially important if computer-linguistic methods supplement free-text search.

A special case in spell checking is verification of names: *Maier* instead of *Meier* is not entered by the user within the meaning of a spelling mistake; he (she) does not know the correct spelling. In many cases, pronunciation is the link between the different spelling variants. The user often only remembers the sound of the name. It is therefore necessary to provide an additional computer-linguistic function which first transforms the entered term into a phonetic transcription from which all possible graphematic terms are generated (see, for example, Regensburg Phonology: Hitzenberger, 1987).

g) Standardization of spelling

Especially in the case of German texts, databases often contain standard deviations from Duden spelling. These deviations lead to errors if the user is utilizing the "correct" spelling during his (her) search.

• Umlauts or ablauts replaced by vowel + e: *mueßig. müßig*

• Capitalization of words written with a small letter at the start of the sentence

• *ß* stored as *ss*

Capitals as highlight in the text and mixed spelling:

DATEV: Datev; GmbH: GMBH

• Special without blank character: *Herr/Frau* instead of *Herr / Frau*

In the near future, attention will also have to be paid to the implementation of the moderate spelling reform of German which will probably be adopted in 1995.

h) Stop word lists / negative lists

Traditional stop words are all functional words in a language (a*nd, the, a, if,* etc.) which themselves do not have any semantic meaning, but define the relationships between terms. During automatic free-text retrieval excluding computer-linguistic methods, they therefore become "meaningless" because the relation system in the text is deliberately regarded and deleted as irrelevant information (cf. Ruge, 1994b).

Stop word lists normally reduce the descriptor list by around 50%. However, they do not affect the quality of research since users do not select them as search terms, only as processing advantages. A problem occurs in that they often do not contain uncovered homonyms (e.g. *Nur die Firma NUR).*

Additional computer-linguistic algorithms will mainly increase recall, but the different types of word combinations will improve precision.

In the simplest (and most frequently implemented) case, computer-linguistic methods work via an isolated symbol chain. This is generally possible in all cases except during word combination and with hyphenated words. However, the reduction in basic forms can be improved by means of context-sensitive heuristics (partial parsing).

- Syntactic level: *Ehe er ....*: the noun *Ehe* is not a stop word as *er* follows.
- Statistical probability: *Trotz vieler Vorkommnisse*: *vieler* is only used very rarely after the noun *Trotz*
- Semantic homography: *Das Geld auf der Bank*: *Geld* excludes the meaning of *Bank as somewhere to sit*.

Multi-word groups are generated in some systems (see, for example, the Saarbrücken system CTX, Schneider, 1987) by means of complete syntactic analyses which can also be interpreted as precise linguistic definition of context operators (in the same nominal phrase, in the same sentence, etc.).

The limit of traditional free-text systems is attained when the algorithms require a generally different type of contents representation. For example, CONTEXT (GMD IPSI, cf. Haenelt, 1994) analyzes the text of the document along syntactic, semantic and pragmatic lines, stores the thus retained complex contents structure as a conceptual network and must analyze the user query using the same method.

The RELATIO/IR system by IBM uses an interesting temporary form (Rahmstorf, 1994). In this system, the thesaurus itself is organized as a controlled, semantic network. The semantic analysis of nominal phrases in the document reorganizes the user interface structure of the search query in a subtree of the thesaurus network. Since all terms are derived from the thesaurus and are therefore used as "controlled vocabulary" in the more

complex form of the thesaurus, new contents cannot be indexed before the thesaurus is extended intellectually.

Summary: there are a whole range of computer-linguistic modules which can be used to improve content analysis of IRS further. Their use in commercial systems now no longer poses a problem for more complex languages such as German.

Efficient algorithms with satisfactory performances are, for example, offered by SOFTEX (cf. Zimmermann, 1993) or LINGSOFT (GERTWOL product; cf. LINGSOFT, 1994). GERTWOL was the winner of the 1994 MORPHOLYMPICS, a competition organized by the GLDV, in which the computer-linguistic methods being discussed here were subjected to an (informal) comparative test (cf. the special publication of the LDV forum in June 1994).

3.2 Effectiveness of computer-linguistic sub-algorithms

After discussing the general options of integrating computer-linguistic methods, the question remains as to whether the additional time and effort are really worthwhile. If this question is examined, it is noticeable in particular that very clear opinions are normally expressed here. In fact, almost everyone knows in advance what is produced by a retrieval test. Unfortunately, everyone is convinced of the superiority of another system.

This fact gives rise to spurious plausibilities with high superficial persuasiveness, which largely prevent a more in-depth examination of the effectiveness of computer-linguistic components.

3.2.1 Graduated model: additive supplements

The arguments used in research literature and among practicians in favor of computer-linguistic algorithms can best be illustrated by means of a graduated model of linguistic components:

a) Objections are raised against the pure free-text solution because the formal options offered in the retrieval language (for example, truncation and context operators) are insufficient to allow error-free recording of the various word forms of a term. The possibility cannot be ruled out that the user will not think about individual word forms or fully understand the side-effects of truncation (ballast). The attention of the user will also be diverted. He (she) must deal with purely formal considerations, a situation which irritates him (her) in his (her) task of finding the correct content descriptors.

b) In the case of algorithms which are limited to morphological analysis and compound splitting, objections are raised to the effect that they are inadequate. The content terms are related structurally in the document. This becomes most apparent with nominal phrases. For example, a user looking for the complex descriptor *Aufnahmevorrichtung für Kernspinresonanzspektren* does not actually examine documents relating to "Aufnahmevorrichtungen" or which have any connection with "Kernspinresonanz", but documents which connect both terms with the factor *für*, i.e. contains them in a special relation in terms of contents.

c) Objections are raised against syntactically-oriented methods because they only depict structural regularities but do not always agree with content relations. For example, studies of patents in the PADOK-I project (cf. Krause, 1987, Womser-Hacker, 1989) showed that the syntax analysis of the Saarbrücken system CTX only determined 75% of the text relations which agree in terms of contents with the complex descriptor of the search query (cf. Schneider, 1987).

d) Kuhlen argues as follows against the restriction to a)-c):

"The contribution of morphological, syntactic and semantic analysis to the overall performance - "Provision of information" - may be minute. As a result, it is not worthwhile to replace the dominant ad hoc methods - e.g. [...] context operators instead of syntactic parsings [...] by linguistically justified methods." (Kuhlen, 1985:7).

It also corresponds to the general pattern of these arguments that the actual increase in value will result from the component to be supplemented. The content analysis functions of each lower level are classified as not effective enough.

If the linguistic subcomponents of the graduated model (morphology, syntax, semantics and pragmatics) are analyzed as a whole from which no component emerges, there is no sensible reason to agree with the idea behind this chain of arguments, i.e. the thesis that a complete morphological, syntactic, semantic and pragmatic analysis promises the best retrieval results in terms of quality.

Nothing appears intuitively more plausible than to demand that all processes which can be observed in data processing and generation by humans should be automated as far as possible and in a 1:1 relation. At the same time, this is the simplest and most problem-free way to determine the basic design of an information system. There is nothing more undemanding from conceptualization and nothing more complicated than to take individual components from this so understood simulation approach and to justify their selection stringently.

Consequently, there is also no reason not to proceed in this way unless - seen in quite general terms - you do not *want* to (e.g. faced with a 'morally' judging background as in Weizenbaum, 1976) or *cannot*. Both cases mean that - for whatever reason - parts of the individual components shown in the graduated model no longer apply. However, if the complete chain of individual components is not implemented in full, 1:1 simulation of human data processing and generation switches to the type of restricted systems. We are therefore faced with the question of what components these restricted systems should contain and how such a choice can be determined and justified.

With regard to current commercial IRS, it is relatively easy to answer the question of whether all information-linguistic components of the graduated model can be realized. No fully developed systems in a commercial sense are available for the components of semantics and pragmatics. Even experimental systems such as CONTEXT or RELATIO/R (cf. section 3) only cover part of the semantic references contained in a text. Content analysis must therefore by necessity be designed as a restricted system.

The following simple example will illustrate that more computer-linguistic algorithms can not only have no effect in this context, but can also cause further damage (cf. Krause & Womser-Hacker, 1990 for further results of empirical studies).

During the PADOK-I project, the following sequence came about during extensive research tests involving data from the German Patent Office (text basis of the documents: title + abstract, cf. Krause, 1987, Womser-Hacker, 1989) relating to recall (percentage of relevant documents found):

a) Free text never attained the top position although it was clearly favored in the test design.

b) In the overall evaluation, PASSAT (for example, morphology, compound splitting) led to better recall compared with CTX (additionally complex descriptors from a noun phrase analysis). This result was statistically significant in a user group (industry/technical information centers). However, better precision (percentage of ballast) was achieved during retrieval using the database connected to CTX.

An industrial user made the following comments based on his experience with CTX and PASSAT:

"When working with the CTX database, I was irritated by having to think about the complex descriptors during formulation of my own retrieval strategies. The CTX component of the complex descriptors, behind which the additional syntax analysis of CTX stands, meant that I had to consider constantly which text passages were recorded and which were not by the complex descriptors. With PASSAT, however, the basic effect

could be followed relatively easily and incorporated in the cognitive process of formulating a retrieval strategy."

The comments of the industrial user therefore refer to a side-effect of the use of CTX and to additional cognitive stress which might interfere with a potentially positive effect. It is important that this side-effect is not produced through the operation of functionality in the user interface, but is based on the supplied functionality itself. It is not the actual functionality that produces the result, but that what is triggered by this functionality in an overall context.

Consequently, it must always be expected that computer-linguistic components induce processes whose impacts can no longer be controlled analytically (see Krause & Womser-Hacker (1990) for other, even more complex examples). The question of the sequence of systems with different strengths of computer-linguistic components can therefore only be answered empirically. This statement is the specific form of a general information science rule: during a machine-supported information process, single components do not achieve what their function description expresses, but what they bring about in an overall context.

Every change in the text basis therefore calls for new empirical tests. The same applies to a change in other parameters such as user groups, application object, etc.

It is impossible to maintain some widespread spurious plausibilities which assume that an improved (linguistic) performance could at most have no effect, whereby poorer results for richer content analysis methods in terms of computer linguistics would be excluded from the outset. This attitude will at least remain incorrect as long as a complete mechanical simulation of interpersonal cognitive information behavior is not available, for whatever reason. The use of subcomponents, however, leads to the field of restricted systems with their own laws.

4. Statistically oriented methods in information retrieval

Whereas computer-linguistic methods use traditional free-text systems when reducing content analysis, but leave the standard model of IR largely untouched, quantitative-statistical methods (under designations such as best-match or nearest-neighbour methods with or without relevance feedback) change the retrieval process in a more far-reaching manner. They are primarily regarded as techniques against the following negative properties of Boolean retrieval in which user-oriented and cognitive aspects (cf. section 2) are for the most part ignored.

• Boolean retrieval divides the document set - without any interim stage - into two discrete subsets: into documents which fulfill the "exact match" (= relevant documents) and those which do not. Documents with three found terms are rejected in a search query comprising four terms linked by the AND operator in exactly the same way as those with 0.

All outputted documents are equivalent from a system viewpoint. The last document on the results list can best satisfy the information requirements of the user. At the descriptor level, this corresponds to the impulse to use all descriptors as "equally important", something which users regard as an inadmissible simplification.

• Users often have problems in making adequate use of the logical operators AND, OR and NOT. One reason for this is that the semantics of the logical operators do not agree with the semantics of the natural language terms. The priorities of the Boolean operators must also be known.

Commercial developments in statistical systems are only slowly starting to compete with traditional methods on the market. One example is TOPIC (cf. Wood & Moore, 1993 for an overview of commercially available IR software). There are also several experimental system variants which are already being used at universities or scientific institutes for real applications (e.g. the INQUERY system of the University of Massachusetts, Broglio & Croft, 1994).

4.1 Basic common features: ranking of the results list and descriptor sequence

With regard to their theoretical background, non-Boolean retrieval models can be divided into probabilistic (statistical probability theory), vectorial (vector space model) and fuzzy retrieval models (theory of inexact quantities) which interpret the similarity function in different ways. As the TREC studies (Harman, 1993) showed, the theoretical differences have practically no effect on the retrieval results, which is why, in my opinion, it is possible to make use of the basic architecture common to all approaches in application development.

Best-match methods can be characterized by the fact that the user strings descriptors together during the query without using Boolean operators and the most relevant documents should come at the start of the results list. This so-called ranking is generated by the system based on similarity criteria.

4.2 Determination of similarity between query and document

The similarities determined by the system define the ranking of the document on the results list. The most widespread similarity is the so-called "vector dot product" in which the similarity is calculated from the product sum of the (weighting) of the terms which appear in the query and the document. The higher the calculated value, the higher the document on the results list.

Also used are the Cosine measure (normalized, including the document length), the Dice coefficient and the Tanimoto measure (cf. Ruge, 1994a). A minimum similarity is often determined by a certain threshold value or the number of required documents defined by the user is utilized as a limitation criterion.

4.3 Weighting

Weighting of the terms in the documents is normally included in the similarity scale. Weighting is automatically determined for every term in a document in relation to certain quantitative properties of the document or the collection of documents. For example, the number of documents in database n and the frequency of the term t in the document collection are included in the "inverse document frequency" weighting. The equation G=log(N/F(t)) means that general terms (quantitative characteristic: high frequency) contribute less to the relevance of a term than specific terms which seldom occur. The yardstick can also include the frequency of a term in a document (within term frequency). The more often t occurs in a document, the higher its weighting, and the less often it appears in other documents too, the better.

A possible connection with the computer-linguistic approaches is already apparent here. All methods mentioned in section 3 can be defined as pre-determination of t. This ensures, for example, that singular and plural forms are not included as two different terms in the calculations.

Some weighting measures also take account of the number of the different terms within a document and/or specify limits for the occurrence frequency of a term (e.g. t must occur at least three times in the data collection). It is also obvious to introduce weighting rules relating to text types, e.g. to weight terms in headings higher than others. In these variants which have to be determined according to their individual application and whose impact on ranking must be verified empirically, there appears to be great potential for improvement with approaches based on non-Boolean retrieval. Generally speaking, however, this applies both to all quantitative-statistical methods and computer-linguistic algorithms. We must deal with restricted system whose real efficiency is exclusively

empirical and can only be proved in relation to specific user groups and application situations.

Wider use of automatic weighting methods in commercial mass databases is primarily prevented by the fact that the term weightings must be recalculated with every change in the data set.

Weightings cannot only be linked to the document terms. The query term can (also) be weighted. As a rule, the user himself (herself) determines intellectually the weighting which he (she) wants to give his query term in his (her) descriptor list (see, however, relevance feedback).

4.4 Relevance feedback

This method calls for a query with at least two stages and the cooperation of the user. He (she) evaluates the results list by crossing the item in, for example INQUERY if an output document was of "relevance" to him (her). The system therefore knows that it is "correct" and uses this dynamic control knowledge from the current dialog situation to "recalculate" the original query. Query terms, which occur frequently in documents specified as "relevant", are given a higher weighting during this reformulation of the query or they are added to the original query. As a result, the modified results list will better satisfy the information requirements of the user (cf. Robertson & Sparck Jones, 1976).

4.5 Clustering and extension lists

The objective of cluster methods is to divide document sets or their descriptor lists into classes whose members are closely related in terms of contents. The degree of relationship is measured formally by means of the joint occurrence of the descriptors in the document. Various mathematical methods are used in this case (cf. Salton & McGill, 1987). A document cluster therefore contains documents whose related descriptors agree as far as possible with one another. A centroid is determined for every cluster and is regarded almost as the "most typical" representative of the cluster. Descriptor clustering is based on the same basic idea: a descriptor t1 is regarded as closely associated with t2 in terms of contents if it occurs as often as possible together with t1 in the documents in the database. In both cases, the cluster is used to determine relevance. If a document d or a descriptor t is regarded as relevant to the information requirements of the user, the members of its cluster are also regarded as relevant (cf. Sparck Jones & Van Rijsbergen, 1973; Croft, 1980).

Cluster techniques supplement the methods already discussed. Document clustering is hardly used on account of the large amount of time spent on mathematical calculations, especially for updating application-related systems. However, descriptor clustering is used. This includes, for example, the extension list of REALIST (Ruge; 1992; Schwarz & Thurmair 1986). Extension lists are networks of terms which correlate statistically with the starting term. Figure 1 shows a simple example of the term *CPU* and how often (in %) the term *CPU* occurs with other terms in documents (number).

Extension lists give rise to the following working method:

• The descriptors are extracted in an initial step (by means of any IR system such as STAIRS or GOLEM/PASSAT).

• Extension tables are then drawn up based on terms and their occurrence, followed by calculation of the correlations between the descriptors of a document and those of other documents in the data set.

The extension lists may be made available to the user as an additional tool for independent strategy planning (research assistance) or may also act as a basis for automatic research expansion as part of IR components (see, for example, Grefenstette, 1992).

The preparation of extension lists normally does not depend on language. In the REALIST context, extension lists were prepared based on a representative subquantity of the total document set and then transferred to the overall stock.

CPU		in document: 105 correlation terms: 1817	
correlation (%)	Term	correlation (%)	Term
65.71	DATA	7.62	LOGIC
49.52	MEMORY	7.62	LOCATION
33.33	STORE	7.62	COMMUNICA-TION
32.38	CENTRAL	7.62	CHIP
...			

Figure 1: Use of term CPU

4.6 Extended Boolean retrieval

As with computer-linguistic and quantitative-statistical methods, the best-match method and Boolean retrieval also offer a mixed form. The main starting point for this consideration were the empirical observations that both search methods lead to widely differing result quantities and that users of non-Boolean systems regarded the lack of Boolean operators in certain situations as a handicap. They want, for example, to link synonyms by OR. With extended Boolean retrieval, the user starts with a conventional Boolean query which will, however, only determine a preliminary selection of potentially relevant documents from the search strategy. A ranking method is then added to the query in order to arrange the preliminary selection according to actual relevance (cf. Salton et al; 1983, Bookstein, 1981).

In order to ascertain the correctness of the basic conviction that the user's objective - "preliminary selection" - substantially reduces the disadvantages mentioned at the start of section 4, empirical calculations would have to be made for specific application areas. The advantage would have to be so great that it exceeds the higher cognitive burden on the user, which is automatically produced due to concept doubling.

5. Development potential

According to the considerations in sections 1 to 4, content analysis produces three main groups of potential starting points for improving the retrieval performance of current commercial systems based on Boolean algebra which promise to be successful in a relatively short space of time and with limited expenditure.

5.1 Additional computer-linguistic modules

Traditional free-text retrieval reduces content analysis to a symbol-oriented analysis. Documents are regarded as the stringing together of chains of symbols which are separated by blank spaces or punctuation marks. The erroneous matching processes thus produced can be reduced by using computer-linguistic methods. Even if empirical tests are ultimately the only way to find out whether the retrieval performance of a specific application area can be improved substantially by computer-linguistic methods, there are many plus points as regards improved quality.

Computer-linguistic components are now available as fully developed basic software both in the scientific and commercial sectors (examples SOFTEX and GERTWOL).

5.2 Use of quantitative-statistical methods

Various individual components appear highly promising as regards short-term development of current commercial systems based on Boolean algebra.

a) REALIST (Retrieval Aids by Linguistics and Statistics) contains statistical techniques (extension lists) which lead to terms which correlate with a descriptor sought in the database. The user can select or exclude additional terms from the extension list (to prevent an unwanted reference of the selected descriptor) and thus specify his (her) information requirements precisely. This method is well-suited conceptually to linguists and also counters the restriction on Boolean queries, namely that the terms cannot be weighted.

b) Due to the anticipated performance problems with quantitative-statistical methods, tests using an extended Boolean retrieval model appear promising. Thanks to this model, the Boolean query can be carried out using a traditional system such as STAIRS. The thesis is that Boolean logic would no longer be normally deployed by the user for complicated links, but merely to determine a pre-selection. A ranking algorithm, which works with weighted document terms, could then be added to the resulting results quantity. The empirical calculation of specific parameters from the specific application context appears to be a highly promising aspect. As additional "references", these parameters reveal the significance of descriptors (e.g. highlighted position of a term in the title or basic principle of a document).

Based on the information in b), the entire Boolean query could be replaced experimentally in a second development phase by a ranking method with relevance feedback methods. Whether this will lead to improved quality in a specific application and what calculation methods should be selected can only be determined by means of empirical tests using suitable prototypes.

5.3 Summary

The methods discussed in this paper therefore provide a sufficient number of detailed approaches which could improve the information performances of current commercial databases based on Boolean algebra without forcing a complete new start. It would be possible to combine them by integrating additional research strategies, integrating application-oriented components of an intelligent IR system and rearranging the user interface of such a system. The latter aspects had to be excluded from this article due to reasons of

space together with the increasingly urgent problems of connecting text and factual IR systems.

Literature

Bates, M.J. (1989): The design of browsing and berrypicking techniques for the online search interface. *Online Review 13*:407-424.

Biebricher, P., Fuhr, N. & Niewelt, B. (1986): Der AIR-Retrievaltest. In: Lustig, G. (Hrsg.): *Automatische Indexierung zwischen Forschung und Anwendung* (pp. 127-143). Hildesheim: Olms.

Belkin, N.J. (1993): Interaction with Texts: Information Retrieval as Information Seeking Behavior. In: Knorz, F., Krause J. & Womser-Hacker, C. (eds.): *Information Retrieval '93*. Proc. 1st GI-Conference on Information Retrieval. Konstanz: Universitätsverlag.

Bookstein, A. (1981): A comparison of two systems of weighted Boolean retrieval. *Journal of the American Society for Information Science 32*:275-279.

Broglio, J., Callan, J. & Croft, W.B. (1994): INQUERY System Overview. In: *TIPSTER text program phase 1: Proceedings of a Workshop held at Fredericksburg, Virginia* (pp. 47-67). San Francisco: Morgan Kaufmann.

Croft, W.B. (1980): A Model of Cluster Searching Based on Classification. *Information Systems*, Vol. 5, No. 3:189-195.

DATEV 1994: Datenbanken Anwenderhandbuch. Nürnberg: DATEV.

Fuhr, N. (1988): *Probabilistisches Indexing und Retrieval*. Dissertation, TH Darmstadt, Fachbereich Informatik.

Grefenstette, G. (1992): Use of syntactic context to produce term association lists for text retrieval. In: Belkin, Nicholas et al.: *Proceedings of the 15th Annual ACM SIGIR* Conference on Research and Development in Information Retrieval, SIGIR '92 Copenhagen (pp. 89-97). New York: ACM Press.

Haenelt, K. (1994): Das Textanalysesystem KONTEXT. Konzeption und Anwendungsmöglichkeiten. *Sprache und Datenverarbeitung*. Bd. 18, No. 1:17-31.

Harman, D. (Ed.) (1993): *The First Text Retrieval Conference (TREC-1).* Gaithersburg, National Institute of Standards and Technology. Special Publication (pp. 200-207). Springfield: NIST.

Hitzenberger, L. (1987): Phonological Access to Databases. In: Luelsdorff, P. (ed.): *Orthography and Phonology* (pp. 65-76). Amsterdam: Benjamins.

Informationszentrum Sozialwissenschaften (IZ): Jahresbericht 1994. Bonn: IZ.

Knorz, G. (1994): Automatische Indexierung. In: Hennings, R.-D. et al. (Hrsg.): *Wissensrepräsentation und Information-Retrieval* (pp. 138-198). Modellversuch BETID Lehrmaterialien; 3. Potsdam: Universitätsverlag.

Krause, J. (Hrsg.) (1987): *Inhaltserschließung von Massendaten*. Hildesheim: Olms.

Krause, J. (1990): *Zur Architektur von WING: Modellaufbau, Grundtypen der Informationssuche und Integration der Komponenten eines Intelligenten Information Retrieval*. WING-IIR-Arbeitsbericht. Regensburg: Informationswissenschaft.

Krause, J. (1992): Intelligentes Information Retrieval: Rückblick, Bestandsaufnahme und Realisierungschancen. In: Kuhlen, R. (Hrsg.): *Experimentelles und praktisches Information Retrieval* (pp. 35-58). Festschrift für Gerhard Lustig. Konstanz: Universitätsverlag.

Krause, J. & Womser-Hacker, C. (Hrsg.) (1990): *Das Deutsche Patentinformationssystem*. Köln: Heymann.

Kuhlen, R. (1985): Verarbeitung von Daten, Repräsentation von Wissen, Erarbeitung von Information. Primat der Pragmatik bei informationeller Sprachverarbeitung. In: Endres-Niggemeyer, B. & Krause, J. (eds.): *Sprachverarbeitung in Information und Dokumentation* (pp.1-22). Berlin: Springer.

Kuhlen, R. (1991): *Hypertext. Ein nicht-lineares Medium zwischen Buch und Wissensbank*. Berlin: Springer.

LINGSOFT (1994): GERTWOL. Questionnaire for MORPHOLYMPICS 1994. LDV-Forum 11 (1). *Sonderheft MORPHOLYMPICS*: 17-29.

Lustig, G. (1986): Eine anwendungsorientierte Konzeption der automatischen Indexierung. In: Lustig, G. (Hrsg.): *Automatische Indexierung zwischen Forschung und Anwendung* (pp. 1-12). Hildesheim: Olms.

Möller, Tong (1993): *Juris für Juristen*. (Law for Jurists). Dissertation, Universität des Saarlandes.

Rahmstorf, G. (1994): Semantisches Information Retrieval. In: Neubauer, W. (Hrsg.): *Proceedings Deutscher Dokumentartag* 1994 (pp. 237-260). Trier: Universität.

Robertson, S. E. & Sparck Jones, K. (1976): Relevance Weighting of Search Terms. *Journal of the American Society of Information Science*, Vol. 27:129-146.

Ruge, G. (1992): Experiments on linguistically-based term associations. *Information Processing & Management*, Volume 28, No. 3: 317-332.

Ruge, G. (1994a): *Wortbedeutung und Termassoziation. Methoden zur automatischen semantischen Klassifikation*. Dissertation, TU München, Fakultät für Informatik.

Ruge, G. (1994b): *Skript Tutorial Computerlinguistik*. 1st GI-Conference on Information Retrieval, Regensburg, Sept. 1993.

Salton, G. (Ed.) (1971): *The SMART Retrieval System*. Experiments in Automatic Document Processing. Englewood Cliffs: Prentice Hall.

Salton, G., Fox, E. & Wu, H. (1983): Extended Boolean Information Retrieval. *Communications of the Association for Computing Machinery* 26:1022-1036.

Salton, G. & McGill, M.J. (1987): *Information Retrieval. Grundlegendes für Informationswissenschaftler*. Hamburg: McGraw-Hill.

Schneider, C. (1987): Analyse der Texterschließung. In: Krause, J. (Hrsg.): *Inhaltserschließung von Massendaten* (pp. 56-65). Hildesheim: Olms.

Schwarz, Ch. & Thurmair, G. (Hrsg.) (1986): Informationslinguistische Texterschließung. Hildesheim: Olms.

Sparck Jones, K. & Van Rijsbergen, C.J. (1973): A test for the separation of relevant and non-relevant documents in experimental retrieval collections. *Journal of Documentation*, Vol. 29:251-257.

Weizenbaum, J. (1976): *Computer Power and Human Reason*. From judgement to calculation. San Francisco: Freemann.

Wood, Joanna & Moore, Caroline (1993): *European Directory of Text Retrieval Software.* Aldershot: Gower.

Wolf, Gerhard (1992*):* JURIS - Ein denkbarer einfacher Zugang zu allen Informationen, die Sie brauchen*. jur-pc. 4*: 1524-1810.

Womser-Hacker, C. (1989): *Der PADOK-Retrievaltest. Zur Methode und Verwendung statistischer Verfahren bei der Bewertung von Information-Retrieval-Systemen*. Hildesheim: Olms.

Zimmermann, Harald (1993): *Grundlagen und Verwendung der linguistischen Software von Softex.* Saarbrücken: Softex GmbH.

Address:

Professor Dr. Jürgen Krause, Informationszentrum Sozialwissenschaften, Lennéstr. 30, D-53113 Bonn, Germany, Tel.: +49-228/228-1145, Fax: +49-228/228-1120, e-mail: Krause@IZ-Bonn.GESIS.d400.de

CONFERENCE ABSTRACTS

Included in this volume are the abstracts of the papers presented at the conference. All abstracts are printed here as received. The only alteration made was to shorten abstracts to one printed page. They are listed in alphabetical order of the name of the first author of the paper.

To help you identify topics, we have grouped the abstracts by themes below.

Introduction

Text and Computers - Past, Present and Times to Come. Peter Ph. Mohler, ZUMA, Mannheim, Germany

Content Analysis Projects

Global Issues and the New Geo-Politics of Information, Anthony Giffard, School of Communication DS-40, University of Washington, Seattle, U.S.A.

A Dictionary of Typical German Features. Horst-Alfred Heinrich, Institut fuer Politikwissenschaften, Justus-Liebig-Universitaet Giessen, Giessen, Germany

Dynamics of Changes in Russia: Investigation with Open Questions and Content-Analysis (Longitudinal from 1989). Galina Saganenko, Institute of Sociology, Russian Academy of Sciences, St. Petersburg, Russia

Trend Analysis Using Computerized Text Analysis: A Case Study of Transportation News. Jane Torous, Institute of Transportation Studies, University of California, Irvine, U.S.A.

Coverage of Religion at the U.N. Population Conference: A Content Analysis. Maria Verloop, Barbara Waltz, University of Washington, Seattle, U.S.A.

Tools for the Computer Assisted Qualitative Analysis

Integrating Statistical and Text Analysis. Hans Dotzler, Fachhochschule Muenchen, Muenchen, Germany

An Intelligent Multimedia Approach to Automating Qualitative Data Collection and Coding. Jacqueline Haynes, Intelligent Automation Inc., Gary Resnick, Westat Inc., Rockville, U.S.A.

The Computer-Aided Qualitative Analysis of Argumentations and "Leitbilder". Udo Kuckartz, Humboldt Universitaet, Berlin, Germany

Analytic Induction and the Logic of Qualitative Analysis Software. Nigel Fielding, University of Surrey, Ray Lee, Royal Holloway University of London, London, Great Britain

Researching Software for Computer-aided Qualitative Data Analysis: Is "Social Constructivism" a Relevant Theoretical Framework? Wilma C. Mangabeira, Brunel University, London, Great Britain

Computer-Assisted Analysis in Qualitative Social Research - A Comparison. Joerg Struebing, Institut fuer Soziologie, FU Berlin, Berlin, Germany

Writing in Computer-Assisted Qualitative Data Analysis. Seppo Roponen, National Consumer Research Centre, Helsinki, Finland

Speech Recognition and Corpus Analysis. Thomas Wetter, Wissenschaftliches Zentrum, IBM Deutschland, Heidelberg, Germany

Corpus Linguistics

A Lexical Study as Dictionary Adjunct - An Analysis of YET in the Brown Corpus. Nina Devons, Department of English, The Hebrew University of Jerusalem, Jerusalem, Israel

Search Strategies with the on-line Conceptual Database for Medieval German Literature. Klaus M. Schmidt, Department of GREAL, Bowling Green State University/Ohio, Bowling Green, U.S.A.

Literary and Linguistic Analysis

SERAPHIN: A System for the Automatic Extraction of Main Sentences. Jawad Berri, Denis Malrieu, Jean-Luc Mine, CAMS, CNRS/EHESS, Université Paris IV, Dominique LeRoux, EDF-DERIPN/GRETS, Paris, France

Prototype Effects vs. Rarity Effect in Literary Style. Paul A. Fortier, Department of French, Spanish, and Italian, University of Manitoba, Winnipeg, Canada

Multilingual Text Analysis: Contrastive Concordance between Leibniz's *Monadologie* and its Translations. Antonio Lamarra, Ada Russo, Lessico Intellettuale Europeo, Centro di Studio del CNR, Rome, Italy

Text Encoding Initiative. Peter Scherber, Gesellschaft fuer wissenschaftliche Datenverarbeitung, Goettingen, Germany

Dictionary Layers Underlying Electronic Texts. Dusko Vitas, Cvetana Krstev, Gordana Pavlovic-Lazetic, Faculty of Mathematics, University of Belgrade, Belgrade, Serbia

Strategies of Text Analysis

Management of Big Data Files, Disaster or Blessing?, René Moelker, Royal Military Academy, Breda, The Netherlands

HyperJoseph: The Hypertextual Organization, Ephraim Nissan, Hillel Weiss, Abraham Yossef, School of Computing and Information Systems, University of Greenwich, London, Great Britain

Electronically Coding Corporate Justifications for Top Management Compensation Using the Method of Successive Filtrations, Joseph Porac, James B. Wade, Tim Pollock, Department of Business Administration, University of Illinois, Champaign, U.S.A.

Networks

A Computer Content Analysis Approach to Measuring Social Distance in Residential Organizations for Older People. Donald G. McTavish, Department of Sociology, University of Minnesota, Minneapolis, Kenneth Litkowski, CL Research, Gaithersburgh, U.S.A.

Network Approaches to the Analysis of Texts. Roel Popping, Department of Social Science, Information Technology, University of Groningen, Groningen, The Netherlands

Reference Concepts

Defining Key Words and Concepts through Computational Text Analysis, Renata Fox, Pomorski Fakultet U Rijeci, Faculty of Maritime Studies, Rijeka, Croatia

The Enemy Within: Auto-Correlation Bias in Content Analysis and in Historiometry and Scientometry, Robert Hogenraad, Dean McKenzie, Colin Martindale, Psychology Department, Catholic University of Louvain, Louvain,Belgium

SERAPHIN: MAIN SENTENCES EXTRACTION AUTOMATIC SYSTEM

JAWAD BERRI, JEAN-LUC MINEL, DENISE MALRIEU, DOMINIQUE LEROUX

The SERAPHIN Project aims at producing a knowledge based system in order to extract main sentences in scientific and technical texts. The extract obtained by the system intends to provide the user a rapid overview as an aid in deciding whether the full text is worthy to be read. The method uses surface analysis but tries and reaches deep information from linguistic clues present in the text. Domain knowledge is not required; the knowledge-base leans on linguistic and metalinguistic knowledge and on relations between linguistic and graphic systems. This method is developed on the basis of contextual exploration already applied to many areas of linguistic engineering by the CAMS.

Contextual exploration applied to automatic extraction (1) detects position and linguistic clues such as specific words and grammatical markers (tense, aspectual marks) within their textual context, (2) expresses heuristic rules aiming at statuing on the function and the importance of a sentence in the text, using the precedently determined clues. The system is implemented. It deals with an object- representation of the text which maintains its physical hierarchy in sections, sub-sections, paragraphs, sentences, propositions and lexical units (SGML formatted text); it leans on a task arborescence. The first module performs a morphological reduction and cuts out the text into propositions; the second module detects the contextual exploration clues, using organized lists; the last module uses the heuristic rules in order to determine whether a sentence has to be extracted and to resolve eventual conflicts. The system is operational (about 400 clues and a hundred rules).

Berri Jawad, Denise Malrieu, Jean-Luc Minel, CAMS, Centre d'Analyse et de Mathematiques Sociales, CNRS/EHESS/Universite Paris 4, 96 BD Raspail, 75006 Paris, France, e-mail: minel@cams.msh-paris.fr
Dominique LeRoux, EDF/DER, Direction des Etudes et recherches, IPN/FRETS, 1 Avenue du General de Gaulle, 92141 Clamart, France, e-mail: dominique.le-roux@ grets.der.edf.fr

A LEXICAL STUDY AS DICTIONARY ADJUNCT: AN ANALYSIS OF YET IN THE BROWN UNIVERSITY CORPUS

NINA DEVONS

The study is based on a semantic concordance of YET derived by combined manual and electronic processing from a KWIC listing of the 420 occurrences in the "Brown Corpus of Present-Day Edited American English". Following the OED practice with regard to sense discrimination, priority is given to semantic relatedness over part-of-speech identity, when this is conducive to clarity, and comprehension by the reader. In line with the division of the OED entry, three "HOMOGRAPHS" are distinguished: adverbial intensive, temporal adverb and concessive conjunct. The semantic concordance highlights two features which furnish clues to distinguishing between concessive and temporal YET. In front position in a sentence, or preceded by AND, YET is concessive. In final position, or preceded by AS or NOT, YET is temporal.

These two criteria allow discrimination between the two interpretations to be effected mechanically, in 88% of the concessive realizations in the Corpus and 74% of the temporal.From the semantic concordance, statistical data are derived and presented in the form of a table showing the frequency distribution of the three "homographs" discriminated (and of subcategories reflecting grammatical, semantic and stylistic features) over five Domains of Language Use: 1 Press, 2 Literary Non-Fiction, 3 Popular Non-Fiction, 4 Learned Non- Fiction and 5 Fiction. The Concordance itself furnishes the supporting illustrative examples (full coverage but reduced context length) of all the features presented in the table. Points of interest not treated in contemporary monolingual dictionaries are discussed: stylistic variation between final and non-final sentential position of temporal YET; idiomatic use of YET preceded by "be" or "have" and followed by an infinitive; criteria for determining when "AS YET is a more formal variant of YET, and when YET and AS YET have a quite distinct distribution".

Nina Devons, Department of English, Hebrew University of Jerusalem, Mount Scopus, Jerusalem 91905, Israel, e-mail: ninaa@huji.vms.ac.il

INTEGRATING STATISTICAL AND TEXT ANALYSIS

HANS DOTZLER

During the last twenty years, computerized data management and data analysis tools have faciliated both the establishment of large quantitative studies and the application of powerful multivariate statistical analysis procedures in the social sciences. Similarly, the successful development and installation of software for qualitative text analysis is likely to result in voluminous longitudinal or comparative qualitative studies becoming the state of the art in interpretive analysis.

This paper outlines the theory and practice of actual strategies for integrating quantitative and interpretative analysis and proposes a general and flexible approach.

Hans Dotzler, Fachhochschule Muenchen, Fachbereich Sozialwesen, Am Stadtpark 20, 81234 Muenchen, Germany, e-mail: dotzler@sozw.fh-muenchen.de

ANALYTIC INDUCTION AND THE LOGIC OF QUALITATIVE SOFTWARE

NIGEL FIELDING, RAYMOND LEE

During the first period in the development of software to aid qualitative data analysis methodologist and developers argued that software did not pose a threat to established analytic procedures. Packages were represented as being sensitive to prevailing approaches to analysis such as "grounded theory" and "analytic induction". The principal contribution of the software was to be in automatic data management, and the main analytic metaphor was "code-and-retrieve".

However, qualitative software packages have now undergone several years of refinement, and increasingly support procedures, routines and features which are new to qualitative analysis. Furthermore, debate about these procedures increasingly reveals limitations and flaws in the epistemological foundations of qualitative analysis. It is less and less plausible either to argue that the software is merely an aid to code-and-retrieve or to argue that code-and-retrieve is the sina qua non of qualitative analysis.

Our paper will focus on two software-based analytic approaches - "Qualitative Comparative Analysis" proposed by Ragin and a hypothesis-test-feature presented by Hesse-Biber - in order to examine the meta-logic underlying these approaches to qualitative analysis. We will identify problems in the application of these approaches and relate to the critique of the concept of analytic induction.

Nigel Fielding, University of Surrey, Department of Sociology, Guildford, GU2 5XH, England, e-mail: n.fielding@surrey.ac.uk
Raymond Lee, University of London, Royal Holloway and Bedford New College, Department of Social Policy and Social Sciences, Egham Hill, Egham TW20 OEX, England

PROTOTYPE EFFECT VS. RARITY EFFECT IN LITERARY STYLE

PAUL A. FORTIER

Since the romantics, and particularly since the surrealists, the standard theory of literary style has been that the rare word or the striking image is the most literally effective, in the sense of affecting the emotions of the hearer or reader. Experimental evidence in cognitive science reported by Rosch and others, recently synthesized by Lakoff, demonstrates that most semantic categories have prototype effects, in the sense of best exemplars shading off through less effective evocations to marginal ones. Thus the word "dog" is the best exemplar of the category "dog", with "canine" at the super-ordinate level, and "retriever" or "beagle" at the subordinate level being less effective evocations of the category. So, it becomes of interest to determine whether the rarity effect or the prototype effect predominates in literary usage.

Data is drawn from Andre Gide's *l'Immoraliste*; the vocabulary evoking semantic categories (i.e. literary themes) has been developed in the context of an earlier study, which also demonstrated the importance of certain themes in the novel, using the usual discourse of literary criticism. In the present study, four categories are analyzed: Beauty, Ugliness, Light, and Night. The internal structure of these semantic categories is measured on the basis of the number of ten French synonym dictionaries, as well as the number of the twenty-five novels used in elaborating Engwall's *Vocabulaire du roman francais* which contain words evoking a given category. Simple frequency of each word and of the aggregate of words comprising the category are examined as well. Preliminary results suggest that in the case of this novelist at least, the prototype effect is a more important stylistic feature than the rarity effect.

Paul A. Fortier, University of Manitoba, Department of French, Spanish and Italian, Winipeg, Manitoba, R3T 2N2, Canada, e-mail: fortier@ccm.umanitoba.ca

DEFINING KEY WORDS AND CONCEPTS THROUGH COMPUTATIONAL TEXT ANALYSIS

RENATA FOX

Based on a computational analysis of a textual corpus of 120,000 running words taken from newer publications in the field of international management, the paper isolates key socio-concepts which can be used to identify the communication symbols of management as part of its semiotic system.

Departing from a preview of those investigations of language registers which have tended to concentrate on thematic words, invariably neglecting key words and concepts and failing to separate them from other high frequency words, the paper points to the often neglected value of computational text analysis in the field of sociolinguistics. The paper questions too, the scientific validity of arbitrarily compiled lists of words, and challenges a still present attitude about the impossibility to compile a text corpus without subjective criteria.

Results of the analysis show that the lexical structure of management ergolect is multi-layered (thematic words, key words, professionalisms and slang...). The lexical segment playing the most important sociolinguistic and rhetorical role is the layer of high (work, people) and low frequency (money, profit) key words and concepts which function as communicators of the principles of the social group and its culturological and civilisational values.

Renata Fox, Faculty of Maritime Studies, Studentska 2, 51000 Rijeka Hrvatska, Croatia, e-mail: fox@brod.pfri.hr

GLOBAL ISSUES AND THE NEW GEO-POLITICS OF INFORMATION

C. ANTHONY GIFFARD

With the collapse of the former Soviet Union, the traditional alignment of states into East-West camps has become largely obsolete. What is not yet clear is what new geo-political alignments are emerging as part of the new world order.

This study examines international news agency coverage of three vital global issues to identify these changes. The events covered are the Rio Earth Summit (1992), the U.N. Human Rights Conference (1993) and the U.N. Population Conference (1994). The news agencies were selected to represent differing regional perspectives: The Associated Press (United States) Reuters (Europe) and Inter Press Service (a Third World news agency with a southern perspective).

The study takes advantage of TEXTPACK's ability to process large amounts of data, and analyzes the entire output relating to these conferences from each agency over one-month periods -- two weeks prior to each conference, and the conferences themselves. It seeks to determine what nations and regions are represented, who the actors and sources are, and which topics are given prominence.

This material has not been presented or submitted for publication elsewhere, although one paper, dealing with some aspects of the Rio summit, was read at an IAMCR conference.

C. Anthony Giffard, School of Communications DS-40, University of Washington, Seattle WA 98155, U.S.A., e-mail: giffard@u.washington.edu

AN INTELLIGENT MULTIMEDIA APPROACH TO AUTOMATING QUALITATIVE DATA COLLECTION AND CODING

JACQUELINE HAYNES, GARY RESNICK

In this paper, we describe a field-tested, prototype system to automate the administration and coding of qualitative measures. To test the feasibility and applicability of this approach for survey measurement, the prototype was developed around a semi-projective measure assessing an adolescent social development construct. The Automated Separation Anxiety Test prototype consists of three linked systems: the Interviewer's Assistant, the Editor's Assistant, and the Scorer's Assistant. The Interviewer's Assistant uses an "intelligent" multimedia approach and a graphical user interface to control and standardize the sequence of interview events, to control digital audio recording and to capture the content of the interview for subsequent expert system analyses. The Editor's Assistant uses a similar interface to allow an editor to replay the interview, to describe interview content in greater detail as "events", and to enhance and correct the event record. The Scorer's Assistant is an expert system that relies on pattern recognition strategies and production rules using the sound files, synchronous timing files, and event files collected by the Interviewer's Assistant and the Editor's Assistant to arrive at final codes and classifications. Findings from a field test using twenty-nine interviews that were administered, transcribed and coded both conventionally and using the automated system revealed a high level of inter-judge reliability between the expert, a trained novice (human) and the AI expert system, with little difference between the reliability of the AI system compared to the trained novice. The ASAT approach demonstrates the benefits of collecting, maintaining and analyzing data in their "native" form, as speech, movement, text, or numbers, using a seamless integration between media and AI.

Jacqueline Haynes, Intelligent Automation Inc., 2 Research Place, Suite 202, Rockville, Maryland 20850, U.S.A., e-mail: haynesj1@westat.com
Gary Resnick, Westat Inc., 1650 Research Blvd., Rockville, Maryland 20850, U.S.A., e-mail: resnicg1@westat.com

A DICTIONARY ABOUT TYPICAL GERMAN FEATURES

HORST-ALFRED HEINRICH

Within a panel study about the Germans' national identity 1,300 persons were asked for typical German features by an open-ended question. The answers have been digitalized and collected in a dictionary. It contains more than 500 different concepts and is distinguished almost only by one-word-answers. Therefore, it can simply be used by computer-assisted content analysis. Most frequent statements are "Fleiss", "Ordnung", "Sauberkeit".

The advantage of the open-ended form is shown by the comparison between it and a closed-end question asked parallelly. The alphanumeric data yield a different type of answer with regard to the contents. Since the respondents has been requested to mention three typical features the content analysis proves the preference of specific combinations of mentions.

The following transformation of alphanumeric data into numerical codes and their statistical processing together with quantitative data gives evidence of the degree of the respondents' stereotype or differential ingroup perception. Hence, the results can theoretically be linked with H. Tajfel's concept of social identity.

Horst-Alfred Heinrich, Institut fuer Politikwissenschaften, Universitaet Giessen, Karl-Gloeckner-Str. 21E, D-35394 Giessen, Germany, e-mail: horst-alfred.heinrich@sowi.uni.giessen.de

HISTORIOMETRY AND SCIENTOMETRY

ROBERT HOGENRAAD, DEAN MCKENZIE, COLIN MARTINDALE

Many content-analytic studies involving temporal data are probably biased by a horse-doctor dose of auto-correlations. The effect of such auto-correlations is to inflate or deflate the real differences that may exist between the different parts of text being compared. The solution consists in removing the effects due to auto-correlations, even if the latter are not statistically significant. Procedures such as Crosbie's (1993) ITSACORR or McKenzie, Hawkins, Hogenraad, and Martindale's VERTEX neutralize the effect of at least first-order auto-correlations and can be used with small samples. Examples of content-analytic studies with and without removing auto-correlations are discussed.

Robert Hogenraad, Dean McKenzie, Colin Martindale, The CABAL group, Catholic University of Louvain, Psychology Department, 20 Voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium, e-mail: hogenraad@upso.ucl.ac.be

THE COMPUTER-AIDED QUALITATIVE ANALYSIS OF ARGUMENTATIONS AND LEITBILDER

UDO KUCKARTZ

The analysis of argumentations and "leitbilder" (models of thinking) is an upcoming field of qualitative data analysis that requires specific analytic strategies. By using examples from empirical research about technology assessment the paper gives first an overview about methodological issues of leitbild analysis by focusing on the question: What are argumentation and "leitbilder" and which dimensions and functions do they have?

In the second part of the paper strategy of computer-aided analysis of leitbilder is presented and discussed. Using the WINMAX 2.0 program examples for the coding of different dimensions and the combination of subdimensions with thematic categories will be given. Unlike in grounded theory methodology where coding is regarded as a distinct and definite procedure the coding of leitbilder always contains an element of uncertainty, since leitbilder are present in an given text passage only to a certain degree. A strategy of implementing fuzzy set theory is presented to cope with this problem.

Udo Kuckartz, Humboldt-Universitaet, Institut fuer Rehabilitations-Wissenschaften, Unter den Linden 9/13, 10099 Berlin, Germany

MULTILINGUAL TEXT ANALYSIS: CONTRASTIVE CONCORDANCE BETWEEN LEIBNIZ'S MONADOLOGIE AND HIS TRANSLATIONS

ANTONIO LAMARRA, ADA RUSSO

The paper is about the analysis of three versions in different languages of Leibniz's Monadologie. Two of the three texts, the Latin and the German one, are eighteenth-century translations of the original French text which, really, appeared in printing only after them; the analysis of this texts is part of a project about Leibniz's scientific and philosophical works and his terminology.

The three texts have been subdivided into little meaningful contexts (pericopes), every-one accompanied by lemmatized words present in it. Every context and his lemmas is then related to the equivalent one in the other language. The aim of this elaboration is the comparation of philosophical terminology in different languages.

The software Hypercard for Macintosh has permitted to develop a series of stacks and scripts for the processing of all these phases of work. Some facilities have made easy to cut the texts into little part and to associate forms and lemmas. The resultant stacks are used for the processing of a contrastive index and/or concordances and for the printing of a synoptic visualization of the three versions of the text.

Antonio Lamarra, Ada Russo, Lessico Intellettuale Europeo, Centro di Studio del CNR, Via Nomentana 118, 00161 Rome, Italy, e-mail: liecnr@itcaspur.caspur.it

RESEARCH SOFTWARE FOR COMPUTER-AIDED QUALITATIVE DATA ANALYSIS: IS "SOCIAL CONSTRUCTIVISM" A RELEVANT THEORETICAL FRAMEWORK?

WILMA C. MANGABEIRA

The paper will report on an ongoing about the design and diffusion of software for computer-aided qualitative data analysis from a "social constructivist" perspective. By drawing from the metaphor of "technology as text", the paper will address issues of software architecture, user configuration and social relationships between the social scientists "developers and the social scientists users".

The paper will conclude that the social constructivist perspectives offers some new and fresh insights into understanding this new technological development. It will also outline future research questions.

Wilma C. Mangabeira, Brunel University, Centre for Research into Innovation, Culture adn Technology, London, England

A Computer Content Analysis Approach to Measuring Social Distance in Residential Organizations for Older People

Donald G. McTavish, Kenneth Litkowski

Computer content analysis provides another approach to measuring aspects of social structure. Different social positions imply somewhat different perspectives and these social perspectives are evident in language. A language-based approach to the measurement of social distance between positions in an organization is described. The approach taken here uses conversational interviews with occupants of positions in nursing homes. Respondents talk about their organizational situation and the verbatim transcript is used as data. Minnesota Contextual Content Analysis (MCCA), a computer content analysis approach, scores social perspectives taken in these texts and computes a social distance measure as a function perspectives expressed in the texts. This approach facilitates an examination of the impact of this social distance on other organizational and n roles across nursing homes suggest consequences of organizational structure and consequences for the meaning residents express about their experience in the nursing home. Content analysis of the way respondents framed their talk permitted a relatively accurate identification of each respondent with a particular nursing home; an aspect of organizational culture measurable from talk by people in organizations. Finally, the structure of differences between nursing homes reveals important facets of organizational structure.

Kenneth Litkowski, CL Research, 20239 Lea Pond Place, Gaithersburg, Maryland 20879, U.S.A., e-mail: 71520.307@compuserve.com
Donald G. McTavish, Department of Sociology, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A., e-mail: DGM@maroon.tc.umn.edu

MANAGEMENT OF BIG DATA FILES, DISASTER OR BLESSING?

RENÉ MOELKER

Computer assisted content analyses provides the possibility of analyzing big textual data files. But when a researcher analyzes 5436 personnel advertisements which contain over a quarter million words disasters will happen. Methods to manage big data files are needed to control all things that go wrong with computers, software and data processing. In this paper I will go into some results of the research, but the main focus will lie with the problems I encountered. The research was based on 5436 personnel advertisements from the years 1955, 1965, 1976, 1982, 1987 and 1990. The advertisements were published in Dutch national newspapers. They were analyzed by use of the computer program TEXTPACK.

The main findings concern the changing relationship between task complexity of work and qualifications that were demanded by employers. Over time more and more different tasks are mentioned in advertisements. But also there is a development in qualifications. Higher education is demanded. But also socio-normative qualifications, like "flexibility", "creativity", "loyalty", etc. An investigation of this size creates its own problems. A secretary was busy with typing the data for over a half year (possibilities of scanning the advertisements was not available at the time of data collection). The main frame computer was incredible slow. So a "simple" job like splitting the text into lists of words took more than 3 hours. Whenever the researcher made a mistake, he knew the result 3 hours later. Then he could start all over.

Making dictionaries and tagging was an other problem. This was a time consuming work because of the size of the data. With longitudinal data meanings of words are important. The same word may mean something different 30 years ago. Occupations change.

All kind of requirements, validity and reliability checks, that are normal in content analyses, were abnormal in this research. An intercoder reliability could only be made by taking a sample.

René Moelker, Royal Military Academy, Postbox 90154, 4800 RG Breda, Netherlands

TEXT AND COMPUTERS - PAST, PRESENT AND TIMES TO COME

PETER PH. MOHLER

The past of text and computers is, as it is always the case with the past, full of stories of alphanumeric heroines and heroes and their adventures among the demons of numerical computing. Despite the fact that the theory of computing was developed to foster symbolic manipulations, the big machines, nowadays called mainframes, were claimed by natural scientists and accountants as their property and territory. But, as is also always the case with heroines and heroes, in the end the demons are pushed back to their proper place. Live becomes quiet and easy and the heroines and heroes retire to a good life. However, on first sight the 'good life' resulted in a decay of methodological developments; critics could point to the fact that most new developments were either efforts to port existing mainframe solutions to PCs (like OCP to Mini-OCP or TEXTPACK to TEXTPACK PC) or exercises of programmers to computerize already existing solutions. On the other hand, the number of computer applications for text analysis increased and are now spread over many fields. Similarly, while "machine-readable" texts were once scarce, today we have an abundance of texts and material digitized and ready for computational analysis. This indicates that a evaluation of the present would characterize the state of affairs as a transition phase from experimental applications to standardized scientific routines. The glimpse of the future permitted by the abstracts submitted for the Mannheim conference reveals stunning developments. The possibility of computerized transcription of spoken language, for instance, which would be a major break through for qualitative research, the combination of standard tools with hypertext applications enhancing the possibility to cumulate knowledge about specific texts in a comprehensive fashion, and efforts to strengthen the connection between text analysis and theory point the way towards future potentials. And, more important than these spectacular technological advancements, one can observe progress in the area of combining structural, syntactic, semantic and pragmatic characteristics.

Peter Ph. Mohler, ZUMA, Post Box 122155, 68072 Mannheim, Germany, e-mail: mohler@ZUMA-Mannheim.de

HYPERJOSEPH: THE HYPERTEXTUAL ORGANIZATION

EPHRAIM NISSAN, HILLEL WEISS, ABRAHAM YOSSEF

HyperJoseph is a project in hypermedia and AI-based knowledge representation, applied to the Biblical story of Joseph in Potiphar's house (Genesis 39), to its traditional commentaries and supercommentaries, to the associated legendary tradition and literary or otherwise artistic renditions, etc.

An outline of the whole project was outlined in an AIBI'94 paper, whereas a companion paper at the same conference develops a representation for one very specific topic. In this paper, instead, we focus on the hypertextual structure we developed, and on the various modes of information retrieval we allow. A dense web of pointers has been superposed on the textual material. Elaborations accessed are linguistic, narratological, etc., and are linked to either specific (sub)strings (or sets thereof), or to any of e.g. linguistic features, particular threads of exegesis, motif occurrences, narrative units, the characters, the story itself, and contexts in which the story can meaningfully fit, from either the narratological viewpoint, or the coign of vantage of teleological or otherwise religious hermeneutics.

From the viewpoint of hypertext, access links are implemented by explicit definition. However, the HyperJoseph project is also special in that a more general conception of access is adopted: in general, access can be considered to be the result of query processing. The simplest case is just following predefined links, whereas more complex access should be yielded by selecting such data that match a query, that in turn can be just one step involved in symbolic reasoning as based on common-sense knowledge on social or cultural norms or on genre conventions.

The project is eclectic, but this paper in particular is practical, and is meant to describe and exemplify the system as implemented.

Ephraim Nissan, Hillel Weiss, Abraham Yossef, School of Computing and Information Systems, University of Greenwich, Wellington Street, Woolwich, London SE18 6PF England e-mail: e.nissan@greenwich.ac.uk

NETWORK APPROACHES TO THE ANALYSIS OF TEXTS

ROEL POPPING

In the last years several new approaches to content analysis have been established. Several of them are directed to building networks based on the content of the texts. The Network analysis of Evaluative Texts (NET) approach affords data on latent propositions that can be logically derived from texts' manifest content. The approach affords inferences about how such tacit propositions are related to the social contexts within which texts are authored. Map Extraction, Comparison, and Analysis (MECA) affords data and inferences regarding similarities and differences in the ways that groups of individuals relate (i.e., cognitively map) various aspects of their worlds.

The attempts to give a meaning to the relation between to objects (e.g., there exists a causal relation between objects A and B, or there is just an association between both objects) have just started. One way to approach this problem is by using developments in the field of graph theory. With respect to the so-called knowledge-graphs relations between objects have been developed and elaborated. A knowledge- graph is a graph in which conceptual knowledge is represented. A knowledge-graph is used primarily to represent a scientific text.

It is investigated how the approaches can be combined, and what this will contribute to content analysis where the goal is to perform inferential analysis and content analysis for more qualitative research. Illustrations will be presented.

Roel Popping, Department of Social Science, Information Technology, University of Groningen, Grote Rozenstraat 15, 9712 TG Groningen, Netherlands, e-mail: popping@ ppsw.rug.nl

COMPENSATION USING THE METHOD OF SUCCESSIVE FILTRATIONS

JOSEPH PORAC, JAMES B. WADE, TIM POLLOCK

In 1992, the US government began to require that corporations include in their annual proxy statement to shareholders a note explaining their executive compensation philosophy and to list specific justifications for the amounts paid to their senior managers. These statements, by law, must be accurate reflections of the decision process of a company's executive compensation committee. As such, they are a fascinating and important window onto not only executive pay but also the relationship between corporations and their shareholders. We have collected and digitized the 1992 justification statements for 300 of the top company's in the US.

In our presentation, we will focus on our electronic content analysis of the justification statements. We have been using FILTSCOR, a program written by David Fan of the University of Minnesota, to count various kinds of concepts and statements in the text. FILTSCOR is a MS-Windows compatible program that allows the researcher to define higher order concepts and to count the instances of those concepts in a text corpus. Because it separates text with target concepts present from text with concepts absent, FILTSCOR permits the researcher to read the text through a series of "filters" of increasingly more specific coding rules.

The end result is a very detailed mapping of the concept space of a text. We will show how FILTSCOR has permitted us to develop coding hierarchies of justification types and to develop a matrix-like semantic representation of the text corpus we are analyzing. Screen shots, output tables, and diagrams showing the logic of our analysis and coding rules will be presented. If the conference organizers would like, we could also bring a working version of the program for demonstration runs.

Joseph Porac, Department of Business Administration, University of Illinois, 350 Commerce West, 1206 S. Sixth Street, Champaign, Illinois, U.S.A., e-mail: jporac@ ux1.cso.uiuc.edu

WRITING IN COMPUTER-ASSISTED QUALITATIVE DATA ANALYSIS

SEPPO ROPONEN

Writing has several uses in qualitative research. Anthropologists and ethnographers have raised the question of researcher's own texts (e.g. Clifford & Marcus 1986; van Maanen 1988; Atkinson 1992). Researcher's field notes, drafts and ethnographies contain constructions and reconstructions of social life. In grounded theory methodology memo writing is an analytical operation in qualitative data analysis (Strauss 1987; Strauss & Corbin 1990). In the grounded theory approach, memos are used to discover concepts and categories, to build theories, and to summarize text segments and patterns in data. In comparative analysis memos are used for distinguishing categories, dimensions and properties.

Some researchers - for instance those who are analyzing narratives - write summaries of different parts of the narrative. Some others write memos to examine and evaluate their coding system. Many other writing techniques have also been found. Qualitative data analysis courses and workshops pay a lot of attention to the role of writing nowadays, too (Becker 1986).

Advanced qualitative software packages support memo writing in different ways. Though the most versatile programs have an elegant implementation of memo writing, there are still many problems in computer-assisted memoing. In general, user's guides do not contain any descriptions of using memos in real research projects. Manuals advise how to write memos but they do not dwell on the various uses of writing in research. The uses of many summarizing memos, summarizations of those summaries and filtering memos by certain category still remain to be represented as a technical task without a content. Primarily, researchers need theories and methods. They also need qualitative software with versatile analysis features. Besides that, qualitative researchers need hints and experiences of memo writing and well-written manuals with real examples.

Seppo Roponen, National Consumer Research Centre, P.O.Box 5, 00531 Helsinki, Finland, e-mail: seppo.roponen@ktk.kuluttajatalo.mailnet.fi

DYNAMICS OF CHANGES IN RUSSIA: LONGITUDINAL FROM 1989 INVESTIGATION WITH OPEN QUESTIONS AND CONTENT-ANALYSIS

GALINA SAGANENKO

The Russian situation is very changeable in last years. Today nobody can understand it very well - nor inside Russia, nor abroad. Sociologists try to explain some details by multiplicity of public opinion polls, offering very supervisional items for questioning and they are interested only in average opinion of a common man.We also try to clearify some main aspects of the whole situation, and we use more adequative approaches. We have been looking for the social changes regularity by panel investigations. We organize each stage once a year from 1989. We involve rather homogeneous social group of intellectuals of Saint-Petersburg: every time they are 250-300 readers of The National Russian Library, all of them have high education. Sometimes for comparison we involve enterprise workers' social group. We prefer open questions for finding dimensions of the social situations and people's consciousness. These structures of significances in these spheres change cardinally: very few positions are equal to themselves, others transform very much, multiplicity of new positions appear. The structures of significances differ for intellectuals and workers. It is become clear that preferably open questions permit to compare different significant structures for: various years, values of some spheres, various social groups, positive and negative estimations. That is why we use this method as many social parameters are utmost changeable, many social investigative dimensions are impossible "to find up" a priory by any sociologist, even of very high qualification. This investigation gives many methodological ideas about opportunities of sociological knowledge, different methods and open questions' one in particular.

Galina Saganenko, Institut of Sociology, Russian Academy of Sciences/Saint-, Petersburg State Academy of Culture, 7-ya Krasnoarmeiskaya, 14/25, 198052 Saint-Petersburg, Russia, e-mail: kanev@emi.spb.su

THE TEXT ENCODING INITIATIVE AND ITS GUIDELINES FOR ELECTRONIC TEXT ENCODING AND INTERCHANGE

PETER SCHERBER

Themes:

- Objectives and targets of standardizing electronic input of texts.
- The main applications: Re-Using, interchange and transformation of textual resources.
- Overview of TEI activities and the completion of the project.
- The decision towards SGML standard (ISO 8879).
- TEI-DTDs and TEI-specific document classes.
- Overall structure of the TEI Guidelines (P3).
- Access to TEI materials.

Peter Scherber, Gesellschaft fuer wissenschaftliche Datenverarbeitung mbH Goettingen, Am Fassberg , 37077 Goettingen, Germany, e-mail: pscherb@gwdg.de

SEARCH STRATEGIES WITH THE ON-LINE CONCEPTUAL DATABASE FOR MEDIEVAL GERMAN LITERATURE

KLAUS M. SCHMIDT

The Conceptual Database for Medieval German Literature is a joint project between the Christian-Albrechts-Universitaet Kiel/Germany and the Bowling Green State University/Ohio. This project joins together the largest textbank for medieval German literature with the Conceptual Dictionary for Medieval German Epics, both of which have been compiled and collected over a period of more than twenty years. In summer of 1995 the project will be made accessible via Internet to an international community of researchers interested in medieval German language, culture, history, society, and literature. The large database can then be searched from a variety of query viewpoints similar to library catalogue searches or searches in databases for abstracts of journals. The important difference, however, will be that the conceptual database of the project is organized along a complex conceptual system similar to the one that provides the organizational basis for Roget's Thesaurus. The basic organization of the system could also be compared to the yellow pages of a phone book. All the information is based on actual texts. The searches in the conceptual database can retrieve exact frequency of occurrence for the entire textbase, individual texts or a combination of texts not only for individual words and their meaning but also for entire conceptual categories like "Love", "Sexuality", "Law", "Animals", "Weapons", etc. Thus the conceptual database will be the premier research tool of the future for scholars interested in content and the search for motifs and themes. It yields an incomparably greater amount of information than traditional reference works, since the conceptual database allows scholars to grasp the full conceptual scope of a complex work, genre, and period at a glance. Moreover, this tool provides new insights into the development of language, literature, and the social history of the given period.

Klaus M. Schmidt, Department of GREAL, Bowling Green State University, Bowling Green, Ohio 43403, U.S.A., e-mail: bgsuopie.bitnet

COMPUTER-ASSISTED ANALYSIS IN QUALITATIVE SOCIAL RESEARCH - A COMPARISON

JOERG STRUEBING

The purpose of this contribution is to draw a comparison between two types of media for computer-assisted qualitative analysis: First, as an "old-fashioned" media, limits and possibilities of a text-oriented data bank program (LARS) will be reported. Second, this type of tool is contrasted with a "modern" specialized supporting program for qualitative analysis in social sciences (ATLAS/ti). The comparison is based on a set of requirements resulting from both, the "grounded theory approach" of Glaser and Strauss as a key concept of qualitative research processes and from experiences in empirical research work.

A recent qualitative study about working styles of programmers serves as empirical background. In this study 25 long-term interviews with software engineers have been conducted and analyzed. By using a data bank program for managing the data and providing search operations for the emerging structure of codes the comparative analysis has been supported. A partial secondary analysis with ATLAS/ti has been performed to reveal differences in the potential of both edp concepts for social research work. The talk stresses that a computer tool must not only comply with the conditions of methods for data analysis (e.g. qualitative content analysis). Rather it should take the whole research act as a process into account. That is, it has to deal with the linkage between data collection, coding sessions and the modelling of theory. It has to support the various tasks during the research process (coding, linking, comparing, memoing, writing reports). And, it should be flexible enough to assist researches with performing their own work style. Finally, a couple of general considerations is made about the potential use of computers in this domain. Looking forward to future developments the talk presents some suggestions for appropriate edp support of qualitative research processes.

Joerg Struebing, Freie Universitaet Berlin, Institut fuer Soziologie, Babelsberger Str. 14-16, 10715 Berlin, Germany, e-mail: jstrueb@fub46.zedat.fu-berlin.de

TREND ANALYSIS USING COMPUTERIZED TEXT ANALYSIS: A CASE STUDY OF TRANSPORTATION NEWS

JANE G. TOROUS

In 1994 a major earthquake struck Southern California, and there was a flurry of attention in the mass media about the damage, public safety, the cost of repairs, and other issues. The initial media attention focused on road repair and reconstruction, since the earthquake immobilized sections of two major freeways. The coverage continued through time, and began to include stories on alternatives to automotive transportation, like riding local trains and telecommuting. This presentation will focus less on specific transportation results, and more on an examination of the use and comparing different enumeration units, for the purpose of trend analysis. We begin with a review of the communications literature, and describe early efforts to use computerized content analysis for trend-tracking. Then, using the transportation data set as an example, we will examine the results, when different units of enumeration are used. In addition to the analysis by the existing categories, we will show the results of a content analysis in which word-frequencies in the text become the unit of analysis. This particular analysis will allow the categories to "emerge" of their own, based on their prevalence. Factor analysis will be used to amplify the "themes". One approach will content analyze just the headlines of the materials, and compare the results with a manual coding of the same content by two human coders. A third approach will provide trend-data, using the current dictionary, but weighting the results based on exogenous factors like the presence of graphics, page placement, and article length. By comparing and contrasting trends using these different enumeration units, we hope to develop a better understanding of how computerized content analysis can be used to identify and track trends in newspaper/media coverage.

Jane Torous, Institute of Transportation Studies, University of California, Irvine, CA 92717-3600, U.S.A., e-mail: jtorous@translab.its.uci.edu

COVERAGE OF RELIGION AT THE U.N. POPULATION CONFERENCE: A CONTENT ANALYSIS

MARIA VERLOOP, BARBARA WALTZ

This paper examines news agency coverage of religious organizations and views during the United Nations Population Conference in Cairo in the Fall of 1994. The Conference's focus made family planning and abortion issues volatile topics. Although abortion comprised but a small part of the Conference's agenda, the issue became the main focus of debate and news coverage. At the center of this debate two traditional religious opponents, the Muslim and Catholic faiths, became allies joined by their shared belief that abortion is wrong. It was this interesting alliance that made religion a compelling topic for further study.

We used TEXTPACK to perform a content analysis of the three major news agencies: Associated Press (AP); Inter-Press Service (IPS); and Reuters. These agencies were chosen for their different approaches to world news as well as the audiences they serve. Using all three agencies allowed us to compare regional differences in news coverage in addition to providing us with a large data base of material to work with. This analysis would have been unwieldly and unmanageable without the use of TEXTPACK. The content analysis shows major distinctions among the three news agencies in their coverage of the Muslim and Catholic faiths. These differences suggest that the news agencies are very aware of the religious orientation of the audiences they serve. This is evidenced by clear differences in the adjectives and verbs chosen to illustrate the Muslim and Catholic positions. In addition, the analysis revealed a lack of attention to any other religion. This was especially intriguing since the Vatican became a highly quoted source during the Conference although only present as an official observer not as a participant. Other results which we shall elaborate on in our paper include disparities among the news agencies in terms of sources cited and organizations interviewed.

Maria Verloop, Barbara Waltz, School of Communications DS-40, University of Washington, Seattle, WA 98155, U.S.A., e-mail: verloop/waltz@u.washington.edu

DICTIONARY LAYERS UNDERLYING ELECTRONIC TEXTS

DUSKO VITAS, CVETANA KRSTEV, GORDANA PAVLOVIC-LAZETIC

This article introduces the concept of multilayered structured text encoding developed with the intention of enabling the future uses of a particular text, which often can not be foreseen during the encoding phase. Although the encoding of logical text layout using SGML introduces substantial change into the domain of text processing, such an encoding is still unsatisfactory mainly because the graphemic and lexical text structures are encoded as sequences of strings and not as language units . The article discusses (1) the possibility of defining a formal word as sequence of graphemes of particular language rather then character strings, (2) the interaction of such an electronic text (abbr. e-text) with the corresponding extracts from electronic dictionary (abbr. e-dictionary). This approach is illustrated on the electronic edition of one collection of proverbs first published in XIX c that has been ever since the unavoidable source for the establishment of popular and literary Serbian and Croatian norm. However, the collection contains numerous graphemic and lexical variations in comparison with the contemporary Serb(or)Croatian usage. For the preparation of electronic edition the following encoding layers were used: (1) The Document Type Definition based on the TEI recommendations was compiled that describes the overall text structure as well as some specific elements (optional and exchangeable parts of proverbs, anaphoric relations); (2) The word was defined on the graphemic level in order to neutralize certain number of graphemic variations; (3) The extract from e-dictionary is incorporated in e-text which supports text indexing by enabling the recognition of morphological and lexical variations (example); (4) Different equivalence relations were defined for certain lexical layers which enables travelling around the text by paths that have not been explicitly encoded. E-text prepared in this way has open structure suitable for further enhancement (such as adding new information, etc.)

Cvetana Krstev, Gordana Pavlovic-Lazetic, Dusko Vitas, Faculty of mathematics, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia

SPEECH RECOGNITION AND CORPUS ANALYSIS

THOMAS WETTER

Speech recognition is an emerging technology which allows to transforms spoken language immediately into written text. Speech recognition starts by analysing acoustic signals in order to select a small set of hypothetised words from a large overall number - several ten-thousands - of possible word forms. However, since there are words with identical pronunciation but different spelling (homophones), acoustic information alone cannot finally decide about the word meant by the speaker. Therefore - and in order to distiniguish words with only slightly different pronunciation hypothetised words are tested as to whether they suit the context already recognised and the context to follow. Principally, this can be done using grammars. Experimental evidence has, however, proved that such grammars would have to be very complicated and that therefore the expenditure to create grammars would be exorbitant. Nowadays large vocabulary speech recognition products such as the IBM VoiceType Dictation System are therefore based on statistical evaluation of large text corpora rendering estimates for a priori probabilities of the word forms in a vocabulary as well as for sequences of 2 (bigram) and 3 (trigram) words. Apart from logistic problems of collecting corpora - including copyright aspects and the task of storing huge amounts of data - we face problems of quality of the corpora themselves and of validity of decisions based on statistical evaluation of corpora.

Thomas Wetter, IBM Deutschland, Informationssysteme GmbH, Wissenschaftliches Zentrum, PO Box 103068, 69020 Heidelberg, e-mail: twetter@VNET.IBM.COM