The Ulm Textbank Management System: A Tool for Psychotherapy Research

Erhard Mergenthaler and Horst Kächele

1. Introduction

After many years of tape-recording and the production of a large collection of verbatim transcripts, we realized the need to establish a major computerized databank for psychotherapy research and education.[1] The documents stored in the ULM TEXTBANK, as it is known, are primarily a collection of open-ended text databases. The main characteristic of these databases is that they can be continuously expanded, as in, for example, a collection of transcribed brief psychotherapies. It is possible to expand such a database to include every newly undertaken brief psychotherapy without ever reaching a state of completeness or representativeness with regard to this type of text. Completeness or representativeness can only be approximated if, say, in a large set of diagnostic initial interviews, careful attention has been paid to sampling according to a variety of variables such as "sex," "age," "diagnosis," "social class," etc. Examples of closed databases are the Bible (see Parunak 1982), and Freud's collected works.

The degree of completeness exhibited by a database also influences strategies for handling the results of subsequent analyses of these texts. There are two main approaches. In the first, results of all the available analyses are actually stored with the text or in direct relation to it. In the second, texts from the database are processed as needed, according to the wishes of the researcher.

The primarily open databases make it possible to apply any new or old computational form of analysis to any of the texts at any time. Nonetheless, one of the TEXTBANK's goals has been to make available

[1]Supported by the German Research Foundation, Sonderforschungsbereich 129, project B2.

to many different researchers the results of previous analyses of this data, gained with such great effort. Therefore the plan is for the ULM TEXTBANK to store such results directly with the texts. Thus, the textbank management system has been designed to facilitate the following tasks:

1. The input and editing of texts selected according to many different points of view and criteria.

2. Management of an unlimited number of texts on the University of Ulm Computer Center's auxiliary storage.

3. Management of an unlimited amount of information on the texts, their authors, and the related text analyses.

4. Management of an open-ended variety of methods for editing and analyzing stored texts.

5. Providing interfaces to statistical and other user packages.

6. Providing a simple, dialogue-oriented user interface when implementing the tasks in 1 to 5 (above).

In summary, the TEXTBANK management system (TBS) is an information system, designed to administer texts and information about the texts in databases, that makes texts accessible by integrating techniques of linguistic data processing and text processing. It features a uniform user interface that assists in the input, processing, output, and analyses of texts.

2. System Architecture

According to the previous definition, the system's structure includes modules for:

(1) manipulating texts (text module)

(2) analyzing texts (analysis module)

(3) managing the databases (databank module)

The user gains access on a computer terminal via a monitor program, which also coordinates the individual functions of the system.

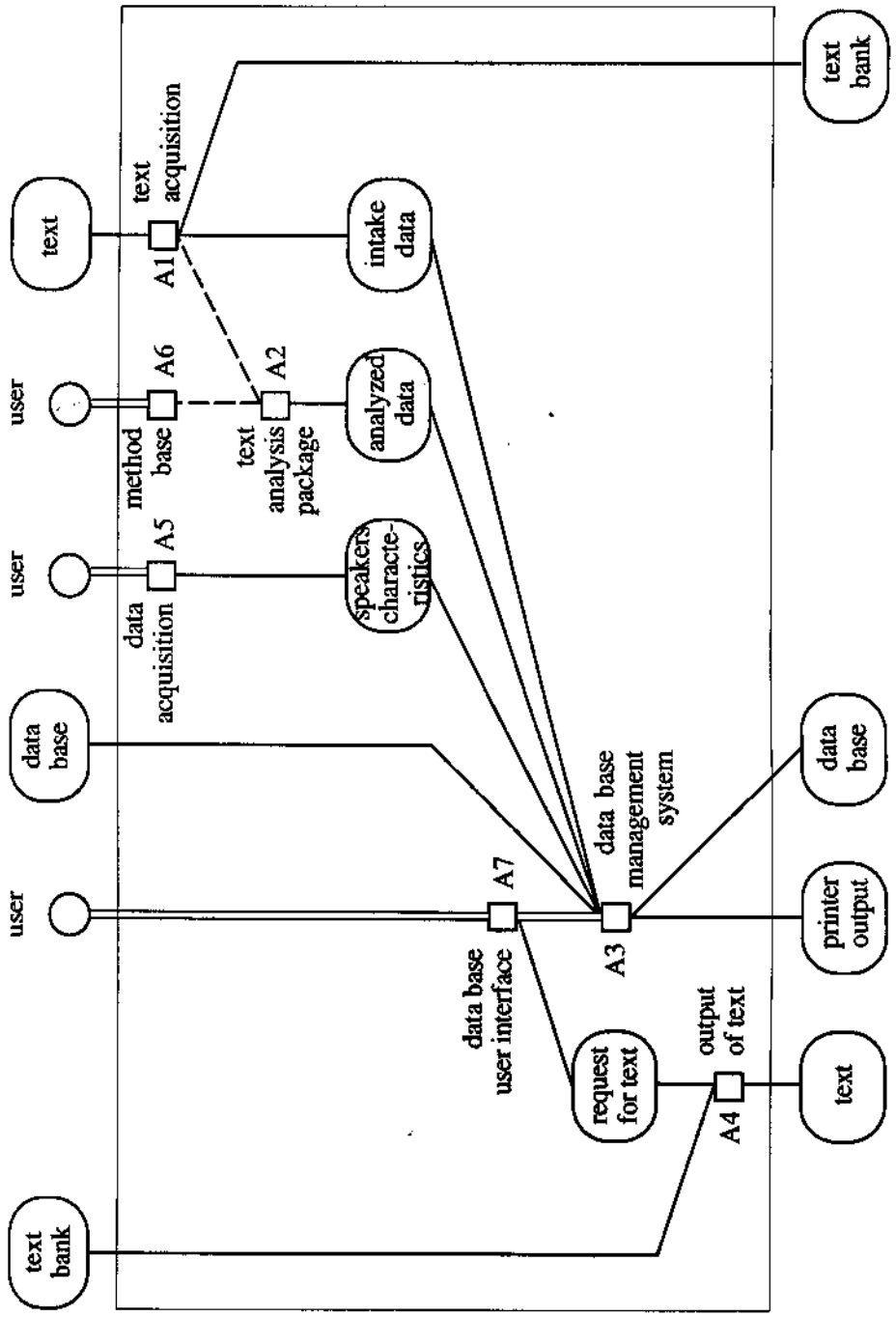
text
text acquisition
A1
intake data
text bank
user
A6
A2
method base
text analysis package
analyzed data
user
A5
data acquisition
speakers characte-ristics
data base
data base management system
data base
user
A7
A3
printer output
data base user interface
request for text
output of text
A4
text
text bank

Figure 1

The Text Module

The text module deals with the input and output of texts in any format (Al and A4 in Figure 1). It is necessary to distinguish between on-line and off-line processing of text. There are two off-line procedures. In the first, a typist produces a computer readable typescript, which is then scanned by a programmable document reader and read into the computer system by the text module. Currently six different typewriter fonts (NLQ) such as COURIER 10 or 12 are supported (see Figure 2).In the second procedure a typist produces an IBM-DOS ASCII text file on a 360 KB diskette which can also be read into the computer system by the text module. In the on-line procedure a typist enters data directly at a computer terminal or personal computer connected to the local area network. In this task the typist is assisted by a special program that deals, for example, with the recognition of typing errors.

The programs of the text module format the incoming text according to the requirements of the system and store it in the TEXTBANK (see Figure 1). At the same time information concerning the speaker and situational factors is entered. When necessary data are missing, these may be obtained and supplemented by the Textbank administrator. These data are stored as "intake data" (together with the labels of the relevant magnetic tape or hard disk) and prepared for further processing.

When output of the text is requested, the system will ask for information concerning the medium on which the requested texts are stored and the nature of the output required. This output may be processed by either the *printout interface* or the *data interface*. Printouts of texts are obtained with the help of special programs that offer the user a variety of options for the output format (see Figure 3). The data interface provides a standard format that is suitable for further linguistic processing within the system.

The Analysis Module The analysis module (A2 and A6 in Figure 1) is a subsystem for the determination of certain characteristics of the texts: (1) formal, (2) grammatical, and (3) content (Table 1). It provides the framework for an expandable body of programs for the analysis of textual data. The module refers the standard text to these programs and stores the results of the analysis as "analyzed data" within the system until they are required for further processing (see Figure 1). Apart from the standard

•12995503008901

•T/you're late, late.

·P/I know, I wouldn't 've, I should have been on time. This is all for my face .

·T/well what happened?

·P /well traffic on Main street was heavy

·T/uh huh

·P/so I came down Irving street and went around Picadilly to the parking lot and I sat there because the parking lot was very full so there were three cars in front of me that had turned the corner, you know, into Picadilly.

·T/yeah

·P/and they had to wait for, what happened was you had to wait for one car to get out and then one went in, the other went out and then one went in. (clears throat) and that was it.

·T/well, so tell me, are you independent professionally? you're still working for uh;

·P / I' m not working for /

·T/really;

·P /they can' t ~fford me .

·T/really finis~leci . okay.

·P/I 'm really finished and I 'm so busy I don't know to do first .

·T/why?

Figure 2

TEXT: 129055 EINHEIT: 0089

1 T: You're late, late.

2 P: I know. I wouldn't 've, I should have been on time. This is all for my face.

3 T: Well what happened?

4 P: Well traffic on *17 street was heavy

5 T: Uh huh

6 P: So I came down *19 street and went around *83 to the parking lot and I sat there because the parking lot was very full so there were three cars in front of me that had turned the corner, you know, into *83.

7 T: Yeah

8 P: And they had to wait for, what happened was you had to wait for one car to get out and then one went in, the other went out and then one went in. (Clears throat) and that was it.

9 T: Well, so tell me, are you independent professionally? you're still working for uh;

10 P: I'm not working for /

11 T: Really;

12 P: They can't afford me.

13 T: Really finished. Okay.

14 P: I'm really finished and I'm so busy I don't know what to do first.

Figure 3

analyses (text size, information content, redundancy, distribution of word categories, and type-token ratio), which are carried out on any text, the user can decide on additional analyses (e.g. the distribution of anxiety themes). Kächele and Mergenthaler (1983) provided an overview of text analyses carried out up to that date. Some examples of text analyses are given later in this chapter.

Table 1

characteristics	monadic	polyadic
formal	text size	speaker sequence
grammatical quotient	verb/adjective	lexical correspondence
substantive	anxiety themes	thematic continuity

The Databank Module

The databank module is a subsystem for the management of information about texts (A3 and A7 in Figure 1). It consists of programs for: (1) the definition of data, (2) data input, and (3) data analysis.

The *definition of data* clarifies the terms used in the description of texts and in the identification of relationships among these characteristics. A standard description scheme can be changed or expanded at any time, thereby controlling the consistency of the data. This program functions as a user interface in a data definition language. The *storage*, *elimination* and *modification* of information that can be described in this scheme is carried out by the programs for *data input*, which also interface with the user in a data manipulation language. During the data input phase the program checks to see if the restrictions specified in the scheme have been observed.

Queries are dealt with by programs for *data analysis*. The interface allows a user to ask questions related to the scheme in a data query

language. The answer is returned to the terminal screen immediately and, if desired, forwarded within the system to the text module.

Associated Demographic data (A5 in Figure 1) are determined by a specific mask-oriented program which facilitates the protection of personal data. The structure of this and related data stored within the system is determined by the data manipulation language described above.

3. Methods and Applications

Starting from a semiotic view of language, which goes back to Peirce, the founder of semiotics, and to its further development by Morris, language is understood as a system of symbols whose structure is determined according to rules based on the relationship between form and content.

Accordingly, it is possible to distinguish between *formal*, *grammatical*, and *substantive* measurements. Each of these types of measurements can be further subdivided with regard to whether it can be applied to a speaker's text or to the entire speech activity in a conversation, i.e., to the dialogue. It is therefore possible to speak of monadic or dyadic values. It is also possible to distinguish among these types of measurements according to the kind of data they utilize. Best known are simple frequencies of occurrence, which form the basis for ratios and distributions.

It should also be noted that, according to the distinction made here, some of the approaches for formal and grammatical measurements presume substantive knowledge, either wholly or in part, for example, the denotative meaning of a word. The contrast with substantive measurements stems from the fact that the required knowledge does not come from the research field itself, namely psychoanalysis, but from the realm of methodologies, i.e., linguistics or information science.

The formal measurements can generally be determined in a simple manner. In computer-aided approaches only the capacity to segment a sequence of symbols (letters, numbers, and special symbols) into words and punctuation is necessary. The programming task is minimal; hardly any recoding is necessary. Table 2 contains a selection of such formal measurements, accompanied by indications of their applicability.

Table 2

Text size (Tokens)	Activity
Vocabulary (Types)	Diversity
Type/Token Ratio	Efficiency
Redundancy	Simplicity vs. Complexity
Distance	Variability, Flexibility
Cluster	Fixation, Focus
Filter	Continuity
Change of Speaker	Dynamic, Rigidity

The simplest and most elementary formal measurement is that of the number of words spoken by the analyst and patient. Kächele (1983) found that in a successful psychoanalytic treatment there was no correlation across 130 sessions in the number of words spoken by analyst and patient. In an unsuccessful treatment of the same analyst these word counts were significantly (+.30, N=110) correlated. O'Dell and Winder (1975) also used the text size as a measure of the therapist's activity in order to distinguish therapeutic techniques. They give 7% as the proportion of therapist's speech in analytic therapy and 31% in eclectic psychotherapy. Zimmer and Cowles (1972), in a study of one patient who visited three therapists with different orientations, also pointed out significant differences. Using the same data, Pepinsky (1979) showed that the therapist's form of activity influences the patients to act in a similar way, i.e., the speech activity of the patient conforms to that of the therapist, as if there is "a convergence of the client toward the level of talk manifested by the therapist" (Pepinsky 1979, p. 7).

The redundancy of a text is a measure adopted from information theory. Spence (1968) proposed some important ideas about psychodynamic redundancy, without testing these ideas empirically. In addition, he formulated a series of hypotheses about the course that redun

dancy takes in psychoanalytic treatment. Kächele and Mergenthaler (1984) confirmed one of these hypotheses, namely that the repetitiousness of a patient's speech increased in the course of treatment. The therapist's values, in contrast, remained constant.

The grammatical measures require the researcher to have linguistic knowledge about the language being studied, for example, the grammar of German. The programming and precoding tasks in the computer-aided procedures are considerable. Moreover, many kinds of questions still cannot be correctly processed automatically. An example is lemmatization, which can assign 50% to 95% of all word forms, depending on the kind of text, to the correct lemma. The psychoanalytic interview, a form of speech with the many syntactically deviant forms (such as incomplete words and sentences) that characterize spoken and spontaneous speech, falls at the lower end of this range. Accordingly, there are hardly any computer-aided studies of psychoanalytic texts using grammatical measures. Table 3 lists a selection of such measures.

The distribution by word type was used by Lorenz and Cobb (1975) to differentiate patients with different psychotic illnesses. To mention one result as an example, they determined that neurotics used more verbs but fewer conjunctions than the normal population used for comparison. At least in German, other variables have to be taken into account, as Eisenmann (1973) demonstrated for conjunctions: "The use of particular conjunctions is determined first and foremost by locality, second by sex, third by age, and lastly by language class" (Eisenmann 1973, p. 407).

The dependence of word choice on word type and semantic class was demonstrated by Busemann (1925) in investigations of children's speech. He spoke of an "active" and a "qualitative" style with regard to verbs and adjectives. He showed that these differences in style are only slightly dependent on the subject being discussed and that they belong rather to personality variables. Using a computer-aided approach to the text of a psychoanalytic interview, Mergenthaler (1985) showed that the realization of a word form within the text may definitely depend on the subject matter. However, this microanalytic view does not exclude the possibility that, viewed at a micro level, personality-dependent variables are effective as described by Busemann.

Table 3

Interjections			Noise	
Word types				
	Cognitive			Structure
	Verb		Action,dynamic	
	Noun		Conditions,	static
	Adjective, Adverb		Features,	modal
	Pronoun		Relations	
Sentences	Relative clause		Complexity	
	Yes questions		Support,	confirm
	Questions		Exploration	
Phrases	Nominalphrases			
	Verbalphrases			
	Prepositionalphrases			
Passive forms	Resistance			
Tense	present			
	past			
	future			
Degree of	Simple		General	
description	Composite		vs.	Specific
Ambiguity	pronoun			
	syntactic			
	lexical			
Diminution	Emotion			
and Raising	Affect			
Ratio	Interjections/Text			size
	Verb/Adjective			
	Relative	clause/main		clause
	Indicative/main			clause
	Phrases/Sentences			

The verb-adjective quotient, introduced by Boder (1940), analogous to Busemann's action quotients, was applied by Wirtz and Kächele (1983) to the first interviews of three different therapists. They concluded that this quotient is a differential measure of the therapist's speech style as well as of differences associated with sex and diagnosis.

The significance of personal pronouns for the structuring of object and self-relations in language has been taken up several times. Several studies have been undertaken on speech material from the ULM TEXTBANK (see Mergenthaler 1985, 1986).

Table 4 shows a selection of *substantive* measures. They presume, in addition to the knowledge mentioned above, detailed expert knowledge of a theory with regard to the theory's area of application. Computer-aided procedures are only able to provide approximate results and are limited to narrowly defined constructions. New approaches in information science, especially in the field of artificial intelligence, could achieve a breakthrough in such matters by establishing data bases in conjunction with a system of rules. Two approaches that strongly emphasize the rule components are Clippinger (1977) and Teller and Dahl (1981a).

Table 4

Themes	
Separation, Mutilation, Guilt, Shame	Anxiety themes
Dogmatism	Cognitive Style
Self, Other	Relationship
Positive, Negative	Affect balance
Primitive Concepts	Cognitive Structures
Speech Acts	
Clarification	Technique
Confrontation	
Exploration	
Interpretation	
Theme Association	Patterns of topics
Theme constancy (after change in speakers)	Interaction Synchrony

The most important kind of the quantitative method for substantive measurements has been content analysis. Gottschalk and Gleser (1969) and Gottschalk (1974) presented the scales most widely used in psychotherapy. Koch and Schöfer (1986) have edited a survey of these methods, including a section by Grünzig and Mergenthaler (1986) on computer-aided approaches. Lolas, Mergenthaler, and von Rad (1982) provide a comparison of results using a computer-aided method with those from other methods.

In a pioneering study, Dahl (1972) was able to trace the downhill course of 363 hours during a 2 1/2 year segment of an unsuccessful psychoanalysis and to convincingly categorize 25 sessions as 10 extreme "work" hours, 10 extreme "resistance" hours, and 5 directly in the middle of the range. Using single words derived from the Havard III dictionary categories he was also able to demonstrate word clusters that manifestly appear to reflect oedipal and other unconscious conflicts (Dahl 1974).

Reynes et al. (1984) used the Regressive Imagery Dictionary (RID) to compare this same patient's 10 working hours and 10 resistance hours. The working hours were characterized by increases in the dictionary categories that assessed primary process language and the resistance hours by increases in the secondary process category scores. This agrees with Freud's earlier attribution of defensive functions to the secondary processes (see Bucci, this volume).

Large continuous segments as well as selected sections of treatment transcripts may thus be examined using *computer aided text analysis* as a tool in psychoanalytic process research (Kächele and Mergenthaler 1984). Further progress requires that methods be developed even more extensively, that basic research be conducted, and that techniques from related scientific disciplines, such as information science and linguistics, be integrated. In fact, Teller (this volume) audaciously proposes artificial intelligence as a basic science for psychoanalytic research.

4. Uses and Sources of the TEXTBANK and the Text Databases

The optimal employment of a TBS in psychotherapy research requires that the text databases to be administered be able to answer the kinds of questions that are likely to be posed. The definition of individual text

databases as subunits of the TEXTBANK are thus especially significant. In this regard two major areas of work have crystallized at the ULM TEXTBANK, each corresponding to a different research approach, *longitudinal* and *cross-sectional* studies.

Longitudinal studies concentrate on text from psychoanalytic treatment; their goal is the study of the change process. Because of the large number of hours of a typical psychoanalysis, it is possible to prepare transcripts for only a small number of different cases. Thus studies of variation (of many kinds) during single cases are foremost.

There are, of course, questions that go beyond variation within individual patients or therapists; these are examined using the texts of initial interviews in cross-sectional studies. Focusing on the initial interview means that many different patients with only one interview can be examined, making it possible to study the effects of variables such as sex and diagnosis, for example. Maintained separately are the text databases required for special investigations, such as those of the Balint group research, verbal exchanges during doctors' visits, and verbal interactions during family consultations.

The texts constituting the major focus of work at the ULM TEXTBANK are being systematically increased. The database of psychoanalysis texts now includes extensive excerpts from four psychoanalytic cases. Individual sessions from nine other psychoanalytic therapies are also included. The database of initial interviews includes several hundred different interviews and is referenced according to the *sex* of the patient or therapist and whether the *diagnosis* is *neurosis* or *psychosomatic disturbance*. This body of texts is being enlarged with special attention to the patient variables of *sex*, *diagnosis*, *social class*, and *age*, and the therapist variables of *experience* and *kind of* psychotherapy (de la Parra 1985).

The kind of texts included in the TBS is determined by the goals, questions, and scientific interests of the supporting and other institutions. For the Department of Psychotherapy at Ulm University this means both the establishment of an empirical basis for research in the field of psychotherapy, and for the support of teaching. The latter takes the form of demonstration materials for the education of medical students and the use of verbatim transcripts in the clinical education and supervision of resident physicians, psychologists, etc. (see Thomä and Rosenkötter 1970).

Table 5

Text type		Textbank			Audio/Video		
	P	T	S	P	T	S	
1	Counseling	1	1	1			
2	Short Term Psychotherapy	4	3	14	25	8	547
3	Analytic Psychotherapy	17	11	78	8	6	782
4	Psychoanalysis	22	13	826	22	9	4557
5	Marital Psychotherapy				*2	2	17
6	Family Psychotherapy	*7	7		43		
7	Group Psychotherapy				*3	2	140
9	Group Work	*4	2		4		
11	Behavior Psychotherapy	1	1	1	3	1	9
12	Initial Interview	300	19	308	218	20	272
13	Initial Interview Report	247	12	291			
14	Psychotherapy Case Notes	5	2	23	1	1	58
15	Psychoanalysis Case Notes	2	1	59	4	1	127
18	Balint Group Work	*2		154	*5	3	150
19	Self-experiential Group	*4	2	43			
20	Dreams	36	2	55			
22	Psychological Testing	16	1	128			
23	Catamnestic Interview	24	1	24			
24	TAT (Thematic Apper-ception Test)	73	6	73			
25	"Narrative"	73	6	73			
26	Genetic Counseling	29	4	29			
29	Individual Reports	-	20	20			
30	Scientific Report	-	40	40			
32	Cognitive-behavioral Psychotherapy	1	1	20			
33	Supervision	6	5	16	8	5	19
34	Psychiatric Interview	8	5	8			
Total		882	16	2311	299	58	6678

P = Patient T = Therapist S = Session. * Couple, Family, Group

Two-thirds of the material in the ULM TEXTBANK has come from Ulm. The other third has been supplied as a result of scientific contacts and joint research projects with institutions outside of Ulm. In most cases these contributions were tied to the right to use the TBS services. While these donations were primarily from the narrow field of psychotherapy, the outside users were almost exclusively linguists who did not require services of the TBS other than the provision of recordings and transcripts along with word and line counts. At present there are contacts with about 30 institutes in Germany, 4 in the United States, 2 in Sweden, 2 in Switzerland and 1 in Austria. Table 5 gives an informal summary of the large variety of materials currently stored in the ULM TEXTBANK. All together, the electronically stored texts include a vocabulary of 135,000 German and 20,000 English words and a total of more than 10 million.

5. Representativeness

Questions as to the representativeness of the ULM TEXTBANK are oriented around the goals of research. However, there are general practical limits such as the large number of hours in each treatment.

In selecting the individual interviews to be stored in the TEXTBANK a number of considerations other than practical ones are important. Several of these, in order of importance, are: the numerical balance among the different therapists, diagnostic categories that are relevant to the central subject of research (anxiety), treatments lasting 300 to 500 hours, and the success of the treatment. Other criteria for selection that would be especially relevant for statistical evaluations, such as the sex distribution of patients and therapists and patients' social class, cannot be accomplished due to the small number of cases. Thus the database of psychoanalytic texts at Ulm is only representative with respect to their specific research goals.

6. Protection of Personal Data

When entering a text into the TEXTBANK, all personal names, geographic references, and other personal characteristics are coded by means of a cryptographic procedure or replaced by *pseudonyms*. While the texts which have thus been made virtually anonymous (Schlörer 1978) are processed at the University's computer center, the key data, i.e., all personal data, remain in the microcomputers used exclusively by

the ULM TEXTBANK. This separate storage, as well as extensive controls on retrieval and manipulation, in large measure protect the ULM TEXTBANK against misuse. The personnel working at the TEXTBANK are subject to professional discretion and instructed as to the relevant regulations with regard to the protection of personal data. The ULM TEXTBANK is registered with the Data Protection List of Baden-Württemberg.

7. Availability and Costs

The services of the ULM TEXTBANK are available essentially free of charge to scientific institutions. Charges are made only for the costs of labor-intensive tasks, such as the transcription of tape recordings and, for university institutions, small charges for materials. In return, it is expected that the texts that are added in this way remain in the TEXTBANK, accessible to other scientists.

With regard to text material loaned by the TEXTBANK, it is expected that a copy of the report of any work using this material be supplied to the ULM TEXTBANK. In this way, in addition to the actual texts, an increasing amount of knowledge about the texts from different disciplines can be stored and made available to others. The ULM TEXTBANK is open to all who wish to store their own texts there. The opportunity for routine or individually tailored text analyses, the convenience of text management, and the variety of options for different outputs should serve as encouragement enough to use the services.

For more technical details concerning structure, data, or usage of the ULM TEXTBANK see Mergenthaler (1985).