THE ULM TEXTBANK[1]

Introduction

Extensive verbatim transcribed protocols of psychoanalytic treatments demanded by Luborsky and Spence (1971) have established themselves as an important source of data in psychoanalytical research. From today's point of view it clearly shows that it was overdue - because of the manifold expectations - to develop, for the application in the area of psychotherapy, proper and user-friendly methods intended for the handling of a text corpus. Beyond this, it also became apparent how important it was to develop meaningful methods for the description of such texts; or to learn from linguistic data processing. To solve the problems thirty years ago in Ulm, an interdisciplinary approach was chosen that connects the psychotherapy-related questions with scientific methods of informatics and linguistics.

Historical Summary

Since 1968 the Department of Psychotherapy of the University of Ulm focused on the development of a methodology for psychoanalytic process research. Within this framework producing audio and video recordings of psychoanalytic long-term treatments provided an essential methodological step that inevitably led to a great collection of verbatim-transcripts. In the course of the first decade we realized the necessity of developing a computer-based databank for our research. Thus began, within the "Special Research Collaboration 129" (SFB 129) of the German Research Foundation, the development of the Ulm-Textbank-Management-System. During this

---

[1] Adapted and shortened version based on Mergenthaler & Kächele (1993).

period of development it became further apparent that such a databank[2] would also serve other scientists who are interested in process research and in analyzing of linguistic material. The final design of the system was therefore characterized strongly by the orientation of a manifoldness of users with very differing methodological approaches (Mergenthaler, 1985). With the conclusion of the SFB 129 in the year 1988, this task was completed. Since then the ULM TEXTBANK is a public institution available for psychotherapy research.

General Aims

One of the main goals in the development of the TEXTBANK was to make available linguistic material of psychotherapeutic sessions and also of neighbouring areas, to researchers in order to save time and money for research endeavours that can be conducted with the already accessible material (archival function). A further goal was also to create availability for computer-based text analyses for all the scientists who do not have resources of this kind of their own. A third goal, consisted in to connect the results that were gained in preceding analyses in order to facilitate a rediscovery of text on the basis of already available results. Thus the TEXTBANK-MANAGEMENT-SYSTEM was designed to facilitate the following tasks.

a) Recording and processing of texts under manifold points of view.

b) Management of an unlimited number of text units on various data media.

c) Management of an unlimited amount of information on text units and their authors and their conducted text analyses.

d) Management of an open-ended amount of methods for editing and analysis of stored text units.

e) Support of interfaces for statistical and other user-software.

f) Support of a simple interactive user interface in the utilization of the above, from a) to e) mentioned tasks.

The TEXTBANK-MANAGEMENT-SYSTEM is thereby an information system, which can manage texts and information about texts and integrates processing of linguistic data-processing as well as text processing for the analysis of texts. It features a

---

[2] In the conceptualization, the example of routine services provided by a medical BLOOD BANK were helpful in many lectures to spread the steering idea of the project.

uniform user interface that assists in the input, processing, output and analyses of text units.

The documents stored in the ULM TEXTBANK represent an open collection of texts. The main character of such data collections is that they can be extended continuously. The measure of completeness of a data base however influences the strategies in which the research results concerning these texts are handled. Two approaches can be discerned: in the first, all available data are stored together with the texts itself; in the second, the analyses are being conducted anew, according to need.

The TEXTBANK project provides the realization of tools for informatics in psychotherapy research. Special interest was given to the acceptance and performance of the, at the time, rather new approach. During the phase of the gathering of texts, the field had to be acquainted with a new fact: namely the shared usage of primary data. Soon a rapidly increasing number of colleagues understood our goals and joined generously in contributing to the success by making their data sources available.[3]

Methods

Clientele and Samples

The optimal display of a TEXTBANK-MANAGING-SYSTEM requires to be open for processing scientific questions that are hard to predict at the time of its inception. Therefore it is particularly important that individual text collections can be put together as subdivisions in the TEXTBANK. In this context two important working emphases have crystallized, which at the same time correspond to two different research approaches: longitudinal studies and cross-section studies.

Longitudinal studies concentrate on the materials from psychotherapeutic and psychoanalytic treatments. Their goal is to investigate changes through the therapeutic process. To collect large numbers of tape-recorded psychoanalytic treatments is still a dream. Therefore, the study of single cases and their evaluation concerning the manifold aspects have remained in the foreground.

Naturally there are also questions that can be studied in cross-sectional designs, for example in the initial interview texts. In these studies different

---

[3] We extent our thanks to our many colleagues who have given us their trust and cooperation

populations of patients are examined. In this way it is possible to observe the influence of variables such as for example, sex or diagnosis (Parra et al., 1988). Sometimes it is useful to keep separate text collections for special studies such as studies of Balint-Groups (Rosin, 1989), the linguistic behavior of doctors and patients during the ward-rounds (Westphal & Köhle, 1982), or the linguistic exchange in family therapies (Brunner et al., 1984).

The texts that correspond to the main goals of the ULM TEXTBANK are sampled as potential users come along. Meanwhile the archive contains, besides several completely available short therapies, also extensive samples on four psychoanalytical treatments. The initial interview corpus consists of several hundred different interviews and is balanced in view of the sex of the patient and that of the therapist respectively; and further in the view of the diagnostic differentiation, neurosis, that is psychosomatic disorder. The kinds of texts that are found in the TEXTBANK also determine the goals, questions and scientific interests of the other supporting facilities. The creation of such a publicly available research basis is likewise useful for clinical education and supervision.

At present about two thirds of the stock of the ULM TEXTBANK dates from investigations we have performed in Ulm itself. The other texts were gained as a result of scientific contacts and joint research projects in facilities outside of Ulm. In most cases these texts were handed over with the agreement to be utilized by other users as well. Many users are psychotherapists themselves, the other users predominantly belong to the fields of linguistics and social sciences. Presently there are contacts to about thirty institutes in Germany, four in the United States of America, two in Sweden, two in Switzerland and one in Austria.

Altogether the electronically stored texts comprise 10 Mill of words generating a basic vocabulary of 180,000 different German words Thus the ULM TEXTBANK can also provide statements on the frequency of words in spoken German such as Dahl (1979) has published the "Word frequencies of spoken American English," on the basis of his database on Mrs. C.

Questions concerning the degree of representativeness of the collected materials are rather difficult to answer. We tried to keep an eye to include a variety of therapists, to include different diagnostic categories, and to get hold of shorter and longer treatments. Still, the psychoanalytic corpus in Ulm can be viewed as

representative only for specific questions. The table 6.1 provides an overview of the material at the end of the year 2006.

Instruments

Departing from a semiotic view of language, as can be traced back to Peirce the founder of semiotics, and its further development by Morris, language is understood as a system of symbols whose structure is ascertained through rules concerning the relation between form and content.

Correspondingly it is possible to distinguish between the following text measures:

a)   formal

b)   grammatical and

c)   content

Each of these measuring methods can be differentiated further in view of focusing on an individual speaker or on the text as a whole, as a dialogue. Therefore, one can speak of monadic or dyadic measurement values. Further, it can be differentiated according to the kind of measurement values. Well known are the simple measurements of the frequency of appearance of tokens, which are the basis for proportional data and their distribution. Moreover, one should heed that some of the formal measures requires knowledge of content; for example, the denotative meaning of a word.

The formal measurements can generally be generated in a very simple way. In computer based approaches, simply the ability of segmenting of symbol-sequences (letters, digits and special characters) concerning words and punctuation can be examined. The effort for programming is comparatively small; recoding or pre-coding is not necessary. Formal measurements that are available encompass:

- Text size (Tokens)

- Vocabulary (Types)

- Type-Token-Ratio

- Redundancy and

- Change of speaker in family- and group conversations.

The simplest and most elementary measurement is the number of words spoken by the therapist and the patient. This will be illustrated in chap. 6.3 using the example of the treatment of Amalia X.

Redundancy is a text measure that derives from information theory. Spence (1968) suggested that redundancy would increase during psychodynamic treatment without providing an empirical demonstration. Thus we confirmed one of his hypotheses, namely that the redundancy of the patient (frequent repetition of words) during the treatment of the patient Christian Y increased stepwise. The values of the analyst however, remained constant (Kächele and Mergenthaler, 1984)

Grammatical measures demand linguistic knowledge about the examined language, for example, about the German grammar. The programming and coding effort for computer-based procedures for such measures yet is quite considerable. Until today many questions cannot be performed completely automatically. An example is the lemmatization that is the automatic back-tracing of an inflected word form, to its basic form, which today, depending on the kind of text, has a degree of effect between 50 and 95%. Psychotherapeutic conversation which displays many syntactic deviations (for example, incomplete words and phrases), is typical for spontaneously spoken language and therefore ranges in the lower area. Corresponding, there are few computer-based analyses of psychotherapeutic texts that are based on grammatical measurements (Mergenthaler & Pokorny, 1989). The ULM TEXTBANK provides the following measures:

-        Distribution of word types

-        Diminution and comparison and

-        Interjection

The connection between the choice of a type of word and the semantic class to which it belongs was already shown by Busemann in 1925 in an examination of child language. He spoke of an "active" and "quantitative" style in relation to verbs, respectively to adjectives. He furthermore showed that these stylistic differences are only minimally dependent on the topic of the spoken word and should rather be seen as being personality related. By a computer-based approach, Mergenthaler and Kächele (1985) analyzed a psychoanalytical session of Amalia X and demonstrated that the choice of the type of word definitely depends on the content of the report. However, this micro-analytical view does not exclude the possibility that, viewed by a macro-level, variables of the personality, as they are described by Busemann, can have an influence.

The role of personal pronouns for the structuring of self and object relationships in spoken language was analyzed by Schaumburg (1980) on the four extensively recorded psychoanalytic cases of the ULM TEXTBANK.

Measures of content have been 'types of anxiety' or 'primary/secondary' processes. Measures of content demand additional detailed expertise in terms of the referential content: what concept does a word stand for? Computer based procedures can, in this case, only deliver approximate results and are limited in the frame of narrowly sketched working-models. Convincing examples have been delivered by Dahl´s analysis of working and resistance sessions in the case of Mrs C (Reynes et al., 1984) or Bucci (1997b) description of patterns of Mrs. Cs discourse in "good" and "troubled" hours. Based on a German adaptation of the Harvard III-Psychosocial-Dictionary, Kächele (1976) could demonstrate that linear combinations of content categories and complex clinical concepts, such as positive and negative transference constellations in connection with selected anxiety topics, could be predicted. His results are based on a single case study of the patient Christian Y, in a sample of 55 sessions; correlations between the clinical concepts and the Harvard-III-Dictionary-Categories were amazingly high figuring between .77 and .91.

Large amounts of text, but also selected segments from treatment protocols, can thus be examined with the help of computer based text analysis as a tool in psychoanalytical process research. This will be illustrated in the following contributions.

These approaches however demand that the available methods are further developed; that basic research is furthered and newer techniques of related scientific disciplines, such as informatics and linguistics, are continuously implemented.

Requirements

In order to include a text into the TEXTBANK it is necessary to remove personal names, names of location and otherwise personal features using cryptographic procedures or even to replace them by pseudonyms to keep the text more readable. These personal data are stored on computers, that are exclusively at the disposal to the ULM TEXTBANK management staff. This separate data storing, as well as extensive control mechanisms, protect the ULM TEXTBANK extensively against misuse. The personal of the TEXTBANK is obligated to abide by the rules of the government controlled data protection regulations.

New research fields and directions

Further methodological progress can only be reached by overcoming the weak points of present research techniques. This begins with the process of collecting data, which is still tied to laborious transcript writing; however, in the meantime this can be conducted more efficiently and reliably because of the development of standards (Mergenthaler & Stinson, 1992). In the meantime further steps of qualitative and quantitative text handling have been essentially improved through multimedia approaches by which very comfortable forms of tools are available for archiving, retrieval, analysing and attributing of texts.

The ULM TEXTBANK began in the eighties as a "big science" enterprise in the mainframe computer world. The successful evolution of the PC provides that text analysis systems are by now established in the daily routine of the scientist and offer themselves for defined analyses. However, it remains desirable that individual research groups push ahead for further development. An example for this is the software CM that allows to measure features of the Therapeutic Cycle Model (Mergenthaler 1996)[4]. CM is a text analyzing tool that produces, for a text transcript, a graphical representation of the emotional and cognitive processes taking place during the session (see also section 6.6).

The relation to other research programs

The services of the ULM TEXTBANK are available for other scientific institutes for a small fee. Fees are asked particularly for work intensive tasks such as the transcribing of texts of tape-recorded conversations as well as for material. However, it is expected that texts that find their way into the TEXTBANK in this way are also available for other scientists in the future. In view of the material that is being handed out by the TEXTBANK a copy of the report or the publication made with the help of this material should be given in return. Thereby, in addition to the texts, a growing stock of knowledge by various scientific disciplines about the texts can be stored and made available for others. The ULM TEXTBANK is open to all researchers who want to store their texts there. The possibility of routine or specialized text analyses, the

---

[4] for technical details go to: http://inf.medizin.uni-ulm.de.

simple type of text management or the possibilities for multiple prints are reason enough to utilize these services.

As a final remark there remains, due to recent laws concerning data protection, only text material can be admitted to the TEXTBANK or borrowed from it if it is factually anonymous; so that there is no indication as to the identity of the participating speakers. This is often difficult to do without changing the content or partially even distorting it. When dealing with older material another difficulty arises; the law does not allow asking the patient for consent at a later time for further questions or for the handling by another research team. The emphasis of the ULM TEXTBANK therefore lies now rather in the consulting and cooperation with interested scientists of various disciplines and in the work with material in which the parameters of the recent valid data protection are given. Table 6.1 reflects in its overview, the stock of the ULM TEXTBANK prior to the last decisive change of the data protection law (federal and state).

Typ	Available as	No. of Sessions
1	Consultation:	
	Transcript, Audio & Video	4
	Audio	1
	Video	1
2	Short-Term-Therapy:	
	Transcript, Audio	153
	Transcript, Audio & Video	17
	Transcript only	2
	Audio	584
	Video	314
	No Information	5
3	Analytical Psychotherapy:	
	Transcript, Audio	27
	Transcript, Video	19
	Transcript only	91
	Audio	1484
	No Information	14
4	Psychoanalyses:	
	Transcript, Audio	1023
	Transkript only	214
	Audio	5662
	Video	13
	No Information	58
5	Therapy of Couples:	
	Transcript only	2
	Audio	37
6	Family Therapy:	
	Transcript, Audio	31
	Transcript only	28
	Audio	11
7	Group Therapy:	
	Transscript only	26
	Audio	140
	Video	21
8	Supportive Psychotherapy:	
	Transkript, Audio	1
9	Group Work:	
	Transcript only	3
10	Client-Centered Therapy:	
	Video	3
11	Behavioral Therapy:	
	Transcript, Audio	6
	Audio	32
	Video	1
12	Initial Interview:	
	Transcript, Audio	127
	Transcript, Audio & Video	23
	Transcript, Video	3
	Transcript only	232
	Audio	180
	Audio & Video	19
	Video	73
	No Information	8
13	Initial Interview Report	
	Text, Audio	8
	Text	365
14	Report of Psychotherapy Session	
	Text	19
	Audio	57
15	Report of Psychoanalysis Session	
	Text, Audio	7
	Text	153
	Audio	163

16 Lectures General
Audio   14
No Information   3
18 Balint-Group
Transcript only   53
Audio   89
No Information   3
19 Gestalt therapy
Transcript only   46
Audio   2
20 Dreams
Transcript, Audio   36
Transcript only   91
22 Psychodiagnostics
Transcript, Audio   128
Transcript only   104
Audio   40
23 Follow-up Interview
Transcript, Audio   41
Transcript only   15
Audio   7
Video   7
24 TAT (Thematic Apperception Test)
Transcript only   183
25 Language Sample
Transcript only   74
26 Genetic Consultation
Transcript only   37
28 HIT (Holzmann-Inkblot-Test)
Text   60
29 Psychotherapy Session Report
Text   19
32 Cognitive Behavioral Therapy
Transcript, Audio   20
Audio   19
33 Supervision
Transcript, Audio   16
Audio   5
34 Psychiatric Treatment
Transcript only   24
36 Family Interview
Transcript, Audio   2
Transcript only   47
37 Interactional Psychotherapy
Transcript, Video   28
Transcript only   1
38 Half Standardized Interview
Transcript, Audio   21
Transcript only   5
Audio   44

Table 6.1 The Stock of the Ulm Textbank

(Overview of text units December 31st 2002)

Brunner EJ, Brunner-Wörner R, Odronitz-Dieterle A (1984) Methodische Grundlagen zur Analyse familialer Kommunikation. In: Brunner EJ (Hrsg) Interaktion in der Familie. Springer, Berlin Heidelberg New York Tokyo, S 89-104

Bucci W (1997b) Pattern of discourse in good and troubled hours. Journal of the American Psychoanalytic Association 45: 155-188

Busemann A (1925) Die Sprache der Jugend als Ausdruck der Entwicklungsrhythmik (The language of youth as expression of developmental dynamics). @@, Jena

Dahl H (1979) Word frequencies of spoken American English. Verbatim, Essex

Kächele H (1976) Maschinelle Inhaltsanalyse in der psychoanalytischen Prozessforschung. (Computer based content analysis for psychoanalytic process research). Habilitationsschrift Med. Fakultät, Universität Ulm; second ed. 1988 in PSZ-Verlag, Ulm

Kächele H, Mergenthaler E (1984) Auf dem Wege zur computerunterstützen Textanalyse in der psychotherapeutischen Prozessforschung (On the way to computer-assisted psychotherapeutic process research). In: Baumann U (Hrsg) Psychotherapie: Makro/Mikroperspektive. Verlag für Psychologie Dr. Hogrefe, C J, Göttingen, S 223-239

Luborsky L, Spence DP (1971) Quantitative research on psychoanalytic therapy. In: Bergin A, Garfield S (Hrsg) Handbook of psychotherapy and behavior change. 1st ed. Wiley, New York, S 408-438

Mergenthaler E (1985) Textbank Systems. Computer science applied in the field of psychoanalysis. Springer, Berlin Heidelberg New York

Mergenthaler E (1996) Emotion-abstraction patterns in verbatim protocols: A new way of describing psychotherapeutic processes. Journal of Consulting and Clinical Psychology 64: 1306-1315

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

Mergenthaler E, Kächele H (1993) Locating text archives for psychotherapy research. In: Miller N, Luborsky L, Barber J, Docherty J (Hrsg) Psychodynamic treatment research - A guide for clinical practice. Basic Books, New York, S 54-62

Mergenthaler E, Pokorny D (1989) Die Wortarten-Verteilung: eine linguo-statistische Textanalyse (The distribution of text types: a linguo-statistical text analysis). In: Faulbaum F, Haux R, Jöckel H (Hrsg) SoftStat 89. Fortschritte der Statistik-Software 2. Fischer, Stuttgart., S 512-521

Mergenthaler E, Stinson CH (1992) Transcriptions transcription standards. Psychotherapy Research 2: 125-142

Parra Gdl, Mergenthaler E, Kächele H (1988) Analisis computerizado de la conducta verbal de pacientes y terapeutas en la primera entrevista diagnostica (Computer based analysis of verbal communications of patients and therapists). Acta Psiquiatrica y Psicologica de America Latina 34: 309-320

Reynes R, Martindale C, Dahl H (1984) Lexical differences between working and resistance sessions in psychoanalysis. Journal of Clinical Psychology 40: 733-737

Rosin U (1989) Balint-Gruppen: Konzeption. Forschung. Ergebnisse (Balint-Groups: Concept, Reserach. Findings): 3. Springer, Berlin, Heidelberg, New York

Schaumburg C (1980) Personalpronomina als Indikatoren interpersonaler

Beziehungen (Personal pronouns als indicators of interpersonal relationships). Dr. rer. biol. hum. Abteilung Psychotherapie, Universität Ulm

Spence DP (1968) The processing of meaning in psychotherapy: Some links with psycholinguistics and information theory. Behavioral Science 13: 349-361

Westphale C, Köhle K (1982) Gesprächssituation und Informationsaustausch während der Visite auf einer internistisch-psychosomatischen Krankenstation (Discourse and information exchange during ward rounds). In: Köhle K, Raspe HH (Hrsg) Das Gespräch während der ärztlichen Visite. Empirische Untersuchungen. Urban & Schwarzenberg, München, Wien, Baltimore, S 102-139