# TBMS: Domain Specific Text Management and Lexicon Development*

S. Goeser and E. Mergenthaler
University of Ulm FRG

**Abstract**

The definition of a Text Base Management System is introduced in terms of software engineering. That gives a basis for discussing practical text administration, including questions on corpus properties and appropriate retrieval criteria. Finally, strategies for the derivation of a word data base from an actual TBMS will be discussed.

## 1. Introduction

Textual data are a sort of complex data object that is of growing importance in many applications. Research projects from such different fields as history, law, social sciences, humanities and linguistics but also commercial institutions are dealing with vast quanitities of text. At Ulm University for instance, a machine-readable corpus of spoken language texts has been built up, with the purpose of support for psychotherapeutic process research. The corpus is administered by a **Text Base Management System (TBMS)**, that integrates the functions of archiving, processing and analyzing an arbitrary amount of text (MERGENTHALER 1985).

Several systems satisfying the TBMS definition were conceived independently in the late seventies. THALLER (1983) reports a system CLIO, a TBMS with a highly differentiated data base component and a method base providing computerized content analysis and comfortable editing. LDVLIB (DREWEK and ERNI 1982) is mainly a text analysis package, where data base management and text processing play a subordinate role. A PC-suited TBM-system, ARRAS (SMITH 1984), supports comfortable text inquiry by concordance and index functions, but has no textbase component. Finally there are two TBM-systems for commercial use, MIDOC (KOWARSKI and MICHAUX 1983) and MINDOK (INFODAS 1983) which have a database component and allow extensive processing of text, but no kind of text analysis at all.

## 2. Definition of a TBMS

From the point of view of a TBMS-user, who is supposed to be a non-programming application-field worker, the system is an instrument to take up, to control and/or to analyze his or her individual texts for domain-dependent purposes. Consequently, a system intended to manage a text bank has the following tasks:

1. Input and editing of texts according to numerous points of view.
2. Management of an unlimited number of text units on a suitably sized auxiliary storage device.
3. Management of an unlimited amount of information concerning the text units, their authors, and the related text analyses.
4. Management of an open quantity of methods for editing and analyzing stored text units.
5. Assistance for interfaces to statistical and other user packages.
6. Assistance for a simple, dialogue-oriented user interface when the tasks mentioned in points 1 to 5 are supplied or performed.

The tasks mentioned in point 1 belong to the domain of text processing. The term **text processing system** can be used if sufficient user assistance is provided.

The task of managing an unlimited number of text units on an auxiliary storage (point 2) is the object of long-term data maintenance. The stored sets are grouped in files and put into magnetic storage. If, in addition, it is possible to administer the data sets with the access methods provided by the operating system, then we will speak of a data maintenance system or a **file management system** (e.g., Archive with BS2000 from Siemens and Datatrieve with VMS from Digital Equipement).

The tasks mentioned in point 3 concern the management of a data base including all of its services. These tasks are fulfilled by general **data base management systems**; the classic functions of such systems are the description, handling, take up, manipulation, and retrieval of data. Data structures can be classified as hierarchic, network oriented, or relational.

Point 4 refers to a collection of methods that, given computer assistance in the user's selection of methods, can be termed a method base. Further assistance, such as method documentation and parameter input, is provided by the **method base management system**. Point 5, on interfaces, is a subset of the tasks described by point 4.

All of these tasks are collected in point 6 with regard to the user interface. Thus the TBMS is an integrated overall system consisting of

| | |
|---|---|
| file management system | **FMS** |
| data base management system | **DBMS** |
| text management system | **TMS** |
| method base management system | **MBMS** |

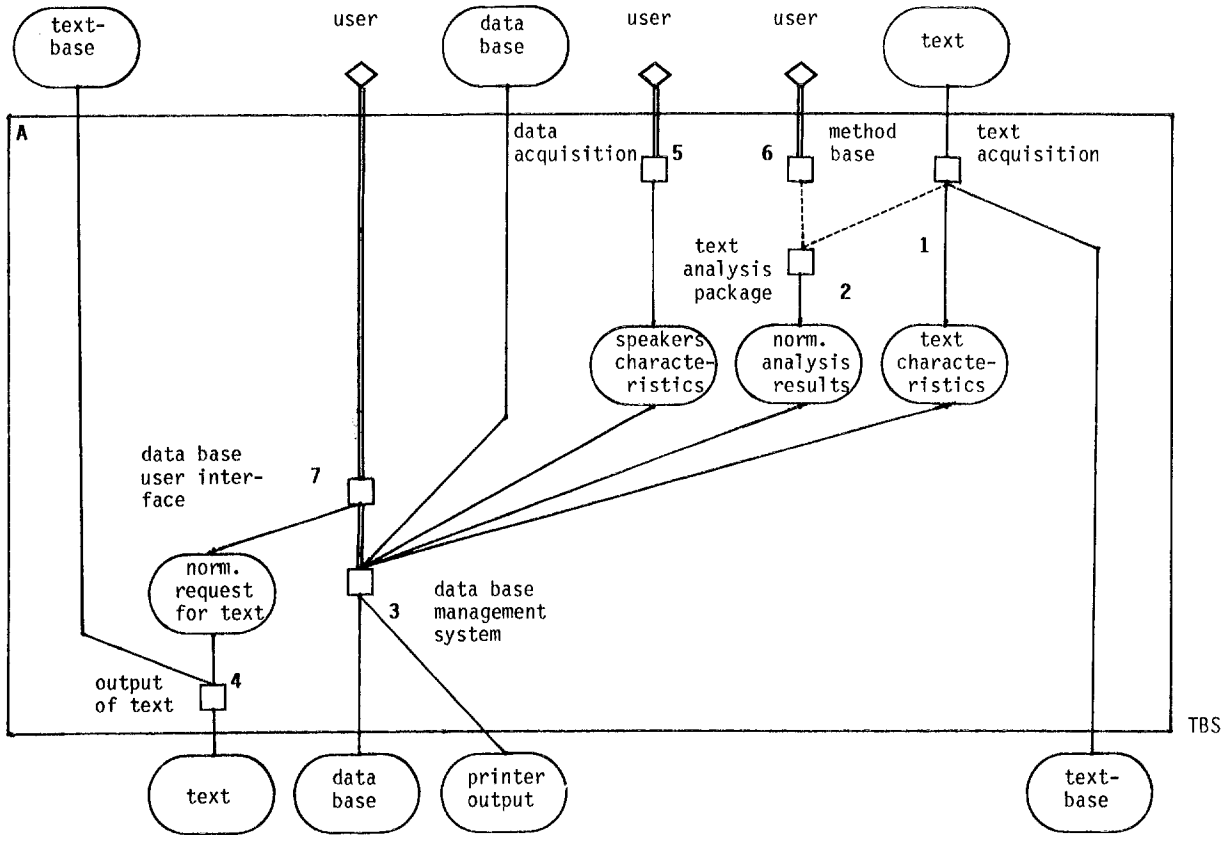


Figure 1

Since the selection of the specific data to be managed by the DBMS and of the methods provided by the MBMS is made in accordance with the kind of texts managed by the FMS, it is legitimate to describe the overall information system as tailormade.

The following definition of the entire system is made, in analogy to the definitions of the individual components, in order to ensure that our terminology adequately reflects this state of affairs:

**Textbase management system - TBMS**
Branch: Linguistic data processing

> The TBMS is an information system that can administer texts and information on texts, and that makes texts accessible by integrating techniques from linguistic data processing and text processing. It features a homogeneous user interface that assists in the take up, processing, output, and analysis of text units.

A System Architecture[1] (top level), which represents one way of realizing a TBMS is shown in Fig. 1.

The TBMS differs from document retrieval systems by containing two additional components: text processing and method base. It is true that the emphasis, as far as retrieval in the TBMS is concerned, is still on data retrieval (for requests made according to the author of a text) and on document retrieval (for text references to be determined according to internal textual features). Fact retrieval is nonetheless an integral component of all the plans for a TBMS, even if systems able to cope with large quantities of material from colloquial speech are still not ready for production. However, this should not be a basis for differentiating within a TBMS. Ideally fact and document retrieval are to be integrated in one system in order to provide satisfactory user assistance.

## 3. Text Base Management

The practical management of texts can roughly be separated into management of texts or text sets (as complex string objects) and of text-related information usually not being part of the text. This information may be pragmatical such as identification and attributes of the text producer(s) or a description of the communicative situation the text originated from, or it may be semantical, specifying some mapping of the text on a formal representation. For both kinds the method base will provide adequate analysis procedures.

A TBMS is indifferent with regard to the completeness of a text corpus, but it supports the more ambitious handling of an **open corpus** in a special way.

The characteristic of an open corpus is that it is an excerpt from a parent population which is continuously being expanded without being tied to the goal of completeness. An example is the collection of fully transcribed short psychotherapies. It is possible to expand this corpus to include every newly undertaken treatment without ever reaching a state of completeness or representativeness with respect to the text type "short psychotherapy". The quality of completeness can only be approximated if there are features limiting the composition of the corpus. Thus a collection of diagnostic first interviews has a higher degree of completeness if it consists of equal parts on the features "sex", "age", "diagnosis", and "social strata". An example of a completely **closed corpus** is the Bible (see PARUNAK 1982).

The degree of completeness of a corpus also influences the strategies for handling the results of text analyses, such as a semantic mapping. There are two principal approaches. In the one, all the available results of analyses are stored completely with the text or in direct relation to the text. In the other, parts of the corpus are processed algorithmically as needed. As detailed by PARUNAK (1982, p.150), users with open corpora tend to use the algorithmic version, while the storage of existing results from previous studies is often preferred for

closed corpora.

Most of the text-related information, be it semantical or pragmatical in nature, can be applied in different ways. First, all of them is clearly of an immediate interest to the system user, who is concerned with a given text. For example, a psychiatrist examining the verbatim protocol of a treatment hour will be interested in its content and in the patient's medical history. He or she might be provided with a weighted semantic category list, a high frequency word list and, as an attribute, the patient's ICD-diagnosis.

Secondly, text-related information can serve within TBMS as an additional criterion (beside text identification and a suitable pre-segmentation) for retrieving texts from FMS. In our TBMS-implementation, the following features do have that retrieval function:

**text type** Given a certain class of destination features a text type can be defined by a set of attribute-value pairs. Intuitivly a text type differentiated that way corresponds to a communicative situation.

**text unit** A text seen as string object can be organized hierarchically, containing (not necessarily continuous) substrings such as utterances or segments, which we call text units.

**speaker** selects among speakers of a given (dialogic) text

**text size** to be specified in current word forms

**theme** on the basis of a semantic category list, textual identification units can be selected according to the weight of a category.[2]

**attributes** For each participant in a communication, a set of feature-value pairs specifying relevant properties is reserved.

## 4. Deriving a Word Data Base

A word data base (WDB) is, formally spoken, a relation ranging over lexicographical features. Every tupel in a WDB specifies a class of feature values being characteristic for a single word form; one or several tupels may be organized in a subset of what has been called a lexical entry (see HESS, BRUSTKERN and LENDERS 1983). Lexical features to be considered here are word form, lemma name, word class, frequency of the word form and some grammatical features, partially depending on the particular word class.

Empirically, a WDB is "based" on one or more communicative situations, that is, it is accumulated with respect to an appropriately sized (see MERGENTHALER 1985) text corpus. Note that situations like, for instance, a psychoanalytic treatment do in fact limit colloquial lexical domains to some degree, simply by imposing thematical restrictions on the text drawn from it.

The relationship between such situation-based domains and the WDB derived from it will be worked out in more detail as follows.

As lexical TBMS-component, a WDB will usually support the method base in analyzing textual properties. We only mention, that most computerized content analysis procedures will operate on raw and lemmatized text, yielding different results. The WDB in our TBMS-implementation (see Table 1) is due to its application in psychotherapy protocol analysis.

The most important procedure in deriving a WDB is automatic lemmatization, defined as providing one pair <lemma name, word class> and optionally inflectional features for every word-form-in-text,

using context for disambiguation. Since TBMS are dealing with mass text, the primary objectives in lemmatization will be first, to have an efficient procedure, second, to have least user assistance and third, to minimize the error rate with respect to the resulting WDB. These objections are tightly linked to a resolution of word class and lemma name ambiguities (homographies) by analyzing context. This seems to be obvious with regard to efficiency and user support. We also stress upon a significant decrease in error rate, in order to avoid interactive lemmatization that goes along with well-known consistency problems.

All current lemmatization procedures (see e.g. KRAUSE and WILLEE 1982) combine **static** components designed for lexicon lookup and **dynamic** components trying to analyze entries not yet contained in the lexicon. Static lemmatization surely will work sufficiently well in all cases where an appropriate lexicon can be provided and a constant domain asserts congruence between this lexicon and the oncoming vocabulary. Dynamic lemmatization, however may become a crucial task in case of non-specific, 'fuzzy' domains like psychoanalytic talks. This is simply a question of the size of vocabulary, which is up to a time not covered by the lexicon. For any given lexicon, this uncovered vocabulary will be more extensive for fuzzy domains than for specific ones.

While unconstrained lexicon access and morphological word analysis are clearly deterministic in nature, context related lemmatization requires a sort of indeterminism on word level. That is, some kinds of lexical pattern[3] should be recognized by the lemmatization algorithm and become stored in the lexicon for later access. This leads us to a matrix description of contextual lemmatization (Table 2).

The dynamic component is applied to those word forms - plus a two sided context of several words - for which no corresponding pattern is available in the WDB. In order to avoid mis-leading lemmatization a linkage between dynamic and static component is built up. A pair <lemma name, word class>, is generated dynamically and will be accepted only if the lemma is element in the WDB. Otherwise, an interactive procedure, the lemmatization dialog, will be initiated.

| | | **procedure** | |
|---|---|---|---|
| | | determi-nistic | nondetermi-nistic |
| **appro-ach** | static | WDB-entry: word | WDB-entry: pattern |
| | dynamic | morpholog. analysis | syntact. analysis |

Table 2

## 5. Final remark

The current implementation of the TBMS ULM TEXTBANK, being finished in 1986, is on a SIEMENS 7.550-D main frame under BS2000 operating system. Further work on the ULM TEXTBANK will include extensions of the method base by robust parsing and rule based content analysis methods.

1) The illustrating technique used here shortly is as follows: Individuals are represented by triangles, dynamic elements as rectangles and static elements as ellipses. The system detail is enclosed in a frame. Controlling operations are indicated by a broken line, reading operations as uninterrupted lines coming from above and writing operations as going to below the frame. Communicating elements are linked with double lines. Lines extending to the frame indicate, that all the dynamic elements contained within it relate to the outside. See Mergenthaler (1985) for a more detailed description.

2) The feature theme is actually not a part of the retrieval system but of the method base, since its

**LEMMA NAME**

1 lemma name
2 variant meanings identification
3 reference to inflected forms
4 semantic description (planned)
5 frequency of occurrence
6 grammatical features (as following)

WORD KIND

| | | | |
|---|---|---|---|
| 1 | noun | 7 | negation |
| 2 | verb | 8 | article |
| 3 | adjective | 9 | preposition |
| 4 | adverb | 10 | conjunction |
| 5 | pronoun | 11 | interjection |
| 6 | numeral | 12 | other |

MORPHOLOGY OF LEMMA

1 simplex (s)
2 compound (c)
3 affixal derivation (a)
4 s, adjective derivation
5 s, verb derivation
6 s, nominal number
7 s, abbreviation
8 s, proper name
9 c, adjective derivation
10 c, verb derivation
11 c, nominal number
12 c, abbreviation
13 c, proper name
14 c, adjective derivation
15 c, verb derivation
16 a, nominal number
17 a, abbreviation
18 a, proper name

| WORD ORIGIN | | GENDER | | PARTICIPLE |
|---|---|---|---|---|
| 1 | foreign | 1 | masculin | 1 have |
| 2 | dialect | 2 | feminine | 2 be |
| | | 3 | neutral | |

**INFLECTED FORMS**

1 inflected form
2 variant meanings identification
3 reference to lemma names
4 semantic description (planned)
5 frequency of occurrence
6 grammatical features (as following)

| RELATIVE FORMS | | NUMBERS |
|---|---|---|
| 1 | diminutive | 1 singular |
| 2 | comparative | 2 plural |
| 3 | superlative | 3 plural only |

| MORPHOLOGY | | CASE/MODE |
|---|---|---|
| 1 | pres.particip. | 1 nominat./indicat. |
| 2 | past particip. | 2 genitive/conjunct. |
| 3 | pres.tense | 3 dative/imperat. |
| 4 | past tense | 4 accusative |
| 5 | infinitive | |

**Table 1**

3) Concerning the problem of differentiating homographs contextually, the straightforward approach of rewrite rules operating on word-class categories won't work except the inventory of those categories is extremely differentiated. For example the rule AD + X + NOUN ---> AD + AD + NOUN is correct only if NOUN is the head of an NP and X is not a present participle or (in English) a noun within that NP. We suppose that a sufficient differentiation may be achieved only with some kind of semantic pattern.

## References


Drewek, R. and Erni, M.: LDVLIB. A (new) Software Package for Text Research. Vortrag ALLC - Conference, Pisa 1982.

Heß, K., Brustkern, J. and Willee, G.: Maschinenlesbare deutsche Wörterbücher. Niemeyer, Tübingen 1983.

INFODAS: MINDOK. Ein Informationssystem auf Kleinrechnern zur Erfassung, Verwaltung und Retrieval von Dokumenten und Daten. (Ausgabe 2.0) INFODAS GmbH, Köln 1983.

Kowarski, I. and Michaux, C.: MIDOC: A Microcomputer System for the Management of Structured Documents. Information Processing 83: 567 - 572 (1983).

Krause, W. and Willee, G.: Lemmatizing German Newspaper Texts with the Aid of an Algorithm. In: Computers and the Humanities 15: 101-113 (1983).

Mergenthaler, E.: Textbank Systems. Computer Science Applied in the Field of Psychoanalysis. Springer: Heidelberg New York (1985).

Parunak, Dyke van, M.: Data Base Design for Biblical Texts. In: Bailey, R.W. (Ed.): Computing in the Humanities: 149-161, North Holland, Amsterdam 1982.

Smith, J.: ARRAS User Manual. State University of North Carolina, Chapel Hill NC, 1984.

Thaller, M.: CLIO: Einführung und Systemüberblick. Manual, Göttingen 1983.



*This work has been supported by the German Research Foundation within the Collaborative Research Program 129, Projekt B2.



Authors address:
S. Goeser, lic.phil and Dr. E. Mergenthaler, Dipl.-Inform.
Sektion Psychoanalytische Methodik
Universität Ulm
Am Hochsträß 8
D7900 ULM