AUTOMATIC INDEXING AND RETRIEVAL AS A TOOL TO IMPROVE INFORMATION AND TECHNOLOGY TRANSFER

Harald H. Zimmermann

September 1982

auch veröffentlicht in:
Proceedings of the IInd scientific meeting COMPUTER PROCESSING OF LINGUISTIC DATA, Bled (Jugoslawien), 7. - 9, Oktober 1982

Abstract

During the last 20 years, linguistic data processing mainly has been seen as a tool to develop linguistic regularities (or detect irregularities) of a given natural language, especially to handle large textual databases ("Corpora"). A second motivation to use a computer was to test some theories or models of a language system (or a part of it) using a simulation program.

As a result of both strategies, the "Saarbrücken Text Analysis System" has been implemented. At present, a very large lexical database is available to analyse written German texts morphologically and syntactically. The syntactic parser is able to handle every German sentence with more than 90% "correct" results.

On the other hand, the development of large (textual) databases within different fields (e.g. law, patent specifications, medicine) is increasing rapidly. Therefore, a computer aided indexing system ("Computergestützte Texterschließung: CTX") has been developed at Regensburg and Saarbrücken University to improve the (even natural language oriented) access to textual data ("free text") applying linguistic strategies to information retrieval processes.

Main results of feasibility studies, especially in the field of German Patent Documentation, are presented.

Introduction

The central problem in this lecture is, first of all, neither a linguistic nor a computer-technical one. The starting point is rather the problem of developing and processing knowledge coded in natural language, in a way which makes it possible for persons and social groups to identify it safely, and to use it for solving their own problems, or for making decisions.

Knowledge coded in natural language is embodied in books, newspaper articles, reviews, legal texts, judgements, reports, patent specifications, journal articles, notices, and so on. In the following, these "texts" will be called "documents" or "document units". By now, about 100.000 periodicals are published all over the world. It is estimated that there are more than 5 million specialized publications (essays, books) in a year /1/. The data base "Chemical Abstracts" covered about 5.5 million document units in 1981, and every year about 500.000 document units are added /2/.

Traditionally, the access to this knowledge is structured by bibliographies, review journals, and collections of file cards. In the meantime, (mostly specialized) classification systems and subject catalogues have been developed. By the aid of such systems, the contents of the documents have been made accessible. By classifying them according to these criteria and rather formal features (for instance author, year of publication, place of publication), a sure retrieval becomes possible.

These works of analyzing texts intellectually involve high staff, they need a highly qualified collaborator; but in the end they are, in many respects, unsatisfactory, because the strong text reduction (normally to few key words) causes a loss of information which hinders the retrieval of relevant documents. Moreover, this procedure requires a specialist of the system with the function of a go-between when making a retrieval. This causes, once again, high cost and complicates the access (time delay).

The information technology presents by now - considered only on the technical side - tools which would allow to connect any interested person by telephone line or by special networks with an information system where the facts are stored electronically (on-line access to data bases).

Usually, the information coded in natural language is, still today, found in printed form which is not directly accessible to computers. Moreover, the possibilities of storage are in spite of considerable improvements of storage capacities still quite restricted. This is why usually only titles or abstracts of essays or books are stored in these data bases (on the intellectual way, so for instance written by the author himself). Other kinds of texts (patent specifications, judgements) are usually represented by the relevant parts of the text. Nevertheless, these quantities are already so huge (the Legal Information System JURIS of the Federal Republic of Germany has already stored more than 200.000 judgements and bibliographical abstracts) that a linguist will feel giddy in the view of the quantity of these "corpora".

Very often the "text part" of such a document in the data base is only an informational instrument - not "understood" by the system -, that is to say: One can only find it and use it in a retrieval process, when the document has already been identified and called up via other items (author, key word ...). Nevertheless, some information systems provide already a so-called "free-text retrieval".

One should suppose that, on this way, linguistic findings are used. Anyway, this is (almost) never the case in the present situation. On the contrary, the method is usually as follows (for instance in the case of the systems DIALOG, STAIRS, and DIRS/GRIPS):

(a) When processing the document, the textual part is diminished by the so-called "stop words" which are stored in a special word list; the remaining words are classified and stored as word forms or chain of characters. A document identification connected with the word form makes sure that a document will be found by a word form of the (free) text.

(b) In the retrieval, that is during the search for information, the word form can be cut off on the "right" side ("truncation", e.g. by using the §-sign), so that the beginning of a word is sufficient to identify all documents containing words with this beginning.

Example:

HAUS§	HAUS	DE 1, DE 2
	HAUSE	DE 5
	HAUSTÜR	DE 23

(c) In the retrieval, free words can be connected by a logical connection (so-called boolean operators such as AND, OR, AND NOT), so that

- synonyms can be considered:

SAMSTAG OR SONNABEND (saturday)

- expressions of two or more words can be found:

JURISTISCH§ UND PERSON§	(text: JURISTISCHE ...PERSON(EN))
LEGAL AND ENTIT§	(text: LEGAL ENTITIES or LEGAL ENTITY)

(d) In most cases, it is possible to connect words already during the development of the system. Thus synonyms can be marked and integrated into the retrieval system.

Two major problems may show that this way of processing free texts in documents is insufficient.

1. Mono-lingual databases involve already the problem of dealing with paraphrases. In some languages - as for instance in English - it may be possible to treat this problem (for words) by methods such as truncation; this is, however, insufficient for languages like German or Dutch, because the composition of words cannot be considered sufficiently. In addition, it is usually the user who has to think about this "paraphrasation" - that means, to make the truncation at the right place or to add synonyms and so on, to say nothing about syntactic paraphrases (German: SCHUTZ VON DATEN eq. DATENSCHUTZ).

2. Databases in foreign languages or multilingual databases have, in addition, the problem of the language barrier (foreign language - native language). Considering the increasing importance of the international exchange of information it is difficult to find attempts for an answer. The microfiche service of INPADOC, for instance, also registers "titles" of Japanese patent specifications: Japanese in Latin transliteration. Even if (in the Western world) English becomes more and more important as an intermediate language (in Japan, a complete abstract service dealing with all important Japanese journals in English is being developed), the task remains to make a connection between the person searching information in his native language and the "foreign language" or "English-speaking" system.

This trend is, at the same time, very dangerous, because very often only those things will be accepted as "relevant" which are written in the communication language of the system - a big barrier in the transfer of knowledge - especially from "smaller" countries - in a utilization all over the world.

Everybody knows that methodical linguistic features are also important in working processes which are not yet mechanized, for instance in the realization of artificial documentation languages, or in the development of terminologies or thesauri. For practical reasons, however, the use of computational linguistic procedures will be, in the following, the main point when discussing the use of linguistic methods in processes of information and documentation.

These procedures can be aimed at:

(1) main developments in the direction of automatic "understanding" of natural language expressions by the computer (AI-systems);

(2) automatic translation of texts/documents (on "any" level) so that either the "human translator" is aided in his work or that the computer (by intellectually developed dictionaries) furnishes sufficiently informative (raw) translations;

(3) improvement of the methods of indexing in order to specify the text processing and to simplify the retrieval.

Considering the complexity of the documents, the first attempt (AI) seems to be the least suitable for furnishing an appreciable easing for the informational practice. It cannot be denied, however, that practically relevant results can be achieved by this method.

By now, the automatic translation has got a chance to be really applied. We refer to the automatic translation of the INSPEC database from English to Japanese (Nagao, Kyoto); the EC uses the system SYSTRAN in some special subjects for making raw translations in order to support the human translators (reduction of cost). The development of a translation system, however, strongly depends on the availability of extensive and highly qualified multilingual dictionaries. This shows that considerable efforts will be necessary before achieving the possibility of producing translations "just as on a production line" (which will not be "perfect" at all anyway).

The procedure of automatic indexing - referred to a natural language - resembles automatic translation - it is even almost the same. Ideally, the procedure consists (as well as the automatic translation) of a linguistic analysis to find the basic forms (morphologic analysis), connection of words (syntactical analysis), semantic disambiguation (semantic analysis), and the construction of so-called "descriptors" in a certain "canonized" form (synthesis). The expenditure, however, is highly reduced compared to automatic translation. This is due to the fact of being monolingual. The component of synthesis is, on the surface, at most a "simple" transformation of deeper structures into an (artificial) language of documentation.

For this reason, procedures of automatic indexing promise the most successful application in practical information services, especially in database systems.

In the Federal Republic of Germany, two main ways in automatic indexing are in the state of research and development:

(1) According to intellectual procedures, few - extremely relevant - keywords are eliminated from texts by automatic methods of statistical and linguistic kind. This procedure was developed at the

technical university of Darmstadt on the basis of English texts. It is being tested on a larger, practical basis in collaboration with the FIZ energy, mathematics, and physics (FIZ 4) at Karlsruhe. First experimental results show that, with this method, one can obtain results which are almost as good as the results of an intellectual indexing /3/.

(2) The way (described in (1)) of concentrating information implies, according to intellectual indexing, considerable loss of information. As an alternative, one has to find methods which

- represent the content of (specialized) texts completely;

- supply results structured in a way that, by suitable retrieval systems according to the application of boolean operators, a concrete search for informations becomes possible.

The research works at the University of the Saarland followed this way. In the following, the procedure will be described as a summary.

2. Computer-aided text processing system (CTX)

The computer-aided text processing system CTX represents the result of many years of research work at the university concerning information software. The models for language analysis developed by basic research work have been further developed - for practical application - into system components for text processing. Building up upon a text processing system for words and sentences, key words (descriptors) with regard to the form and the contents are supplied for German texts/documents.

Legal documents of the subject "data protection" were the base for the first lab-test of CTX. Presently, the system is tested - near to a practical application - in different areas. CTX has a modular structure and interfaces which are independent from the type of the computer. Therefore, it may be integrated into different information retrieval systems or can be added as an indexing component.

For the processing of texts in natural language, CTX has the following tasks:

- The efficient word forms extracted from the text are automatically reduced (by means of a general dictionary with more than 130.000 stems) to their basic form.

  Examples:

text word	basic form
Vorzüge	VORZUG
trat	TRETEN
trifft ... zu	ZUTREFFEN

- In addition, compounds are decomposed into efficient words and disambiguated, if possible.

Examples: Persönlichkeitssphäre
part: PERSÖNLICHKEIT
part: SPHÄRE

- The ambiguity of text words is shown; ambiguous words can be disambiguated (computer-aided) by means of the context.

  Examples:
  ... in der Praxis der Datenverarbeitung ...

  PRAXIS (prakt. Vorgehen)
  (contrary to "Arztpraxis")

- Expressions of two or more words are identified by means of a suitable system of rules.

  Examples:

  tritt ... in Kraft IN KRAFT TRETEN

  personenbezogene, PERSONENBEZOGENE DATEN
  durch das Gesetz
  geschutzte Daten

- Special Simple and Complex descriptors are identified by using suitable dictionaries (which the user can improve and enlarge himself).

  Examples:

  ... modernen Industriestaaten ...

  (simple) INDUSTRIESTAAT
  (complex) MODERNER INDUSTRIESTAAT

- Words without sense ("functional words: THAT, BUT, AND) are eliminated.

The realization of an efficient dictionary component and the (mostly) automatic improvement of the dictionary was of extreme importance. Common words are registered in a morpho-syntactical and in a semantic dictionary. Special words are identified and described in a special dictionary and in a dictionary of semantic relations ("thesaurus"). For special fields, the user himself may determine the structure and the use of words.

The translation system SUSY (developed in Saarbrücken) is the basis of the text processing component; the analysis component of SUSY was integrated in CTX. The linguistic analysis is mainly working on the level of sentence and syntax.

If necessary, the system can be integrated into retrieval components of a special computer. Thus, the analysis and the processing of the description of a problem in natural language during the retrieval becomes possible by using the same rules as for the text analysis (indexing of the documents). This allows a simple formal adaptation of the words used in the question to the indexed words of the texts/documents.

In the following, a short description may explain the procedure of the text exploitation in its components "text analysis" and "supply of descriptors".

INPUT OF THE TEXT

- Input of the (special) text in computer-compatible form (perhaps adaptation to the input-interface of the system). Preparation for the following processing, sentence after sentence.

TEXT ANALYSIS

- Determination of the possible basic forms (if necessary, with decomposition of the compounds) by means of a syntactical dictionary.

- Analysis of syntactical ambiguities.

- Decomposition of the sentence in potential segments (for instance subordinate clauses, co-ordinate sentences).

- Determination and analysis of complex syntactical structures (for instance groups of nouns, verbal groups).

SEMANTIC ANALYSIS

- Reduction of ambiguities of words by means of a system of semantic rules (semantic dictionary).

Result of the text analysis: Text word forms are reduced to basic forms which are partly already definite, the structure of the sentences is determinated.

SUPPLY OF DESCRIPTORS

- Key words, complex expressions, and information concerning their syntactical structures are made available. By means of a specialized dictionary, the ambiguities which have not yet been reduced are analyzed by a specialized weighting procedure.

Result of the supply of descriptors: Descriptors with regard to the form and the contents with additional information concerning the structure, independent of the system.

There are many possible applications for CTX:

- CTX service and CTX databases

Application of CTX as a computer-aided indexing system, used for the construction of databases (information bases);
Application as a procedure for text processing, used for the construction of specialized or non-specialized text-databases;
Application as a means for exploitation of texts within the area of computer-aided determination of contents ("content analysis").

- CTX-IRS (information retrieval system)

Integration of functions of CTX into an information retrieval system.

- CTX-MULTILINGUAL

Retrieval by the integration of foreign-language synonyms into the specialized thesaurus.

- CTX-REGISTER

Construction of registers of basic forms concerning German texts

- CTX/SUSY

The expansion of the CTX-system towards translations (German-English, German-French) using the translation component of the translation system of Saarbrücken is being prepared. The targets of this expansion are

- automatic translation

- documentation and indexing, including the translation of texts or abstracts in special areas.

In the meantime, CTX represents a practicable software system for the solution of problems in the area of exploitation of texts in natural language. At the same time it is a system which is distinguished by high flexibility with regard to the changing requirements of users and a large spectrum of possible applications.

3. Practical applications of the CTX-system

By now, the system has been applied in several large areas. First, at the university an information system was developed and implemented within the area of "data protection". This model served mainly for the clarification of principal procedures, for the construction of a test database (according to the JURIS implementation) and the principal anticipation of potential problems of the users). The database constructed for this purpose covers by now a quantity of more than 150.000 running words /4/.

With the beginning of 1982 the German Patent Office could be interested in a test of the CTX

system. The patent applications and the patent claims are the subject of the information exploitation and documentation, that is of the construction of a prototypical information system. These data - about 40.000 document units per annum - are classified, up to now, on the intellectual way (for instance by means of the International Patent Classification (IPC)). Adding descriptors (with regard to the form and the contents) has the target to improve the access to the documents and, as a consequence, to improve the transfer of knowledge and technology. In the meantime, two different tests have been realized which show at the same time the possibilities (capacities) of CTX:

(A) Test with texts of heterogeneous content

The texts used for this test were chosen "accidentally" by employees of the Patent Office in the beginning of 1982. The documents were taken from different areas (food, chemistry, tools...). Accordingly, the application of the test was confined to the morpho-syntactical part of the system CTX, that is the automatic identification of basic forms and expressions of two or more words, and the elimination of functional words. The semantic disambiguation and the construction of the thesaurus were left out of consideration.

In all, 18 documents were analyzed. The 84 sentences were, in an average, about 24 words long. The 1.954 words (running word forms) are distributed among the word forms as follows:

word forms	1954	100.0%
nouns	553	28.3%
adjectives	132	6.8%
verbs	334	17.1%
functional words/adverbs	935	47.8%

Very important for the efficiency of the system is the proportion of words identified with regard to the volume of the dictionary and the morphological analysis attaching thereto (identification of the potential basic forms/lemmata).

In this connection, it is necessary to make a difference between

(a) completely identified words
(b) words which have been identified by analysis (decomposition and derivation)
(c) insufficiently identified words.

Although the system additionally provides a "pseudo-morphologic" analysis component which allows to integrate "unknown" words into the further analysis without improvement of the dictionary, case (c) does not guarantee a correct lemmatization (identification of the basic forms). Therefore such words are examined intellectually, which means they are integrated into the automatic dictionary before the further processing.

In this test 70 running words or 48 different words were identified only "insufficiently" by the system (3.6% of the text words). This means that the morpho-syntactical dictionary had to be enlarged in 48 cases.

Examples:

ANTIMYOCARD-ARZNEIMITTEL
BROMIERT

247 of the 1.954 word forms (that is 12.6%) were identified by word analysis such as decomposition and derivation. 201 different words (word forms) were decomposed (about 30% of the 793 different words appearing in the text).

Examples:

MOTOR/FAHRRAD
AUSGANG*S/INVERTER

The automatic description produced 1.886 descriptors in all, that is about 100 descriptors per document. 968 Simple descriptors (basic forms) were generated out of the 1.019 running words of the word classes noun, adjective, and verb (auxiliary verbs may be omitted); in addition, 522 descriptors were generated out of the 247 decomposed words.

On the basis of the (noun) structures of the sentences which have been identified by the syntactical analysis, another 396 pre-coordinations (in the terminology of CTX: Complex descriptors) were generated. They are distributed among the word classes as follows:

Type	quantity	example
adjective/noun	106	KOERNIGES GUT
noun KON noun	98	MOTORRAD K MOTORFAHHRAD
noun GEN noun	82	RAHMEN G MOTORFAHRRAD
noun PRP noun	46	AUFHAENGEN P RAHMEN
noun/verb	54	SCHWINGUNG DAEMPFEN
verb MOD verb	10	VERARBEITEN KOENNEN

The whole CPU time (CPU = central processing unit, a mainframe computer TR 440 at the University of the Saarland) was about 602 CPU-seconds used, that is about 5.9 seconds/sentence or 0.25 seconds/word. The morphological analysis needed 166 seconds (27.6%), the generation of the descriptors about 111 seconds (about 18%).

(B) Test with texts of homogenous content

In a second test (summer 1982) a greater and, in addition, thematically more homogeneous quantity of texts concerning the German Patent Office was analyzed and processed. In this case, too, the text was chosen "accidentally" by employees of the Patent Office. This time, however, the documents were taken from the field "cables and wires" which is classified by the IPC number H 01 b. Because of this thematical restriction it became possible to introduce enlarged possibilities of processing and retrieval (semantic and statistic disambiguation, relations in a thesaurus).

The 16 analyzed documents included besides the abstracts also extracts from the complete text. By this, the volume of the text tripled in comparison to the first test (A): test (B) contains 227

sentences with 6,531 words. In an average, the sentences had 28.8 words.

word forms in total	6,531	100.0%
nouns, verbs, adjectives (in total)	3,832	58.7%
functional words	2,699	41.3%

These words are distributed as follows (in contrast to (A), the statistic is based on the different words):

different word forms	1,349	100%
nouns	661	49%
adjectives	404	30%
verbs	284	21%

1,018 word forms (that is 76.8%) were fully identified by the morpho-syntactic dictionary and reduced to their basic forms. In 392 cases (that is 29.1%), a word form was identified by the algorithm for decomposition and derivation; the rest (70, that is 5.2%) required a completion of the dictionary.

As - at the same time - the procedure of disambiguation and the relations in the thesaurus were to be integrated, the system had to be enlarged as follows:

1. Verification of the different words and insertion of ambiguities in a special dictionary of ambiguous words. Specification of the probability of the occurrence of a semantic variant in the thematic area;

2. (Enlargement of the dictionary of semantic rules (as far as the inventory of characteristics and structures allows it);

3. Intellectual enlargement of the thesaurus, with regard to special words of the thematic field;

4. Construction of relations for syntactical derivations between the text words.

The following values were determined for the test material:

- The dictionary of semantic ambiguities had to be enlarged by 26 items. By now, it has 978 items.

Examples:

L	S	DICHTUNG
01	8	isolierendes Zwischenstück: eine Dichtung auswechseln
02	9	das Abdichten: die Dichtung eines Rohres vornehmen

03	0	sprachliches Werk: die Dichtung der Romantik
04	0	Phantasie: Dichtung und Wahrheit

- In total, about 50 "rules" were marked in the dictionary of semantic rules.

Examples:

rule:	6621	X1=99, F1=02, U=01
reading:	6621:	internal sign
	X1=99 :	"find the key-word"
	F1=02 :	"if there is a plural..."
	U=01:	"assign variant number 01".
rule:	6622	X=82, A=01, U=03.04
reading:	6622:	internal sign
	X=82:	"if you find a genitive object..."
	A=01:	"...and if the noun of the genitive object is 'abstract'...
	U=03.04:	"assign variant number 03 or 04"

- 318 words were entered into the thesaurus. Accordingly, 573 relations were generated.

Examples:

AUS SUB	KABEL
UNT SUB	FERNMELDEKABEL
AUS SUB	KABELSEELE
TEI SUB	KABEL
AUS SUB	KUNSTHARZ
OBR SUB	HARZ
AUS SUB	ISOLIERMANTEL
NEB SUB	ISOLIERMITTEL

- The existing derivations were enlarged according to the analyzed text material.

Examples:

VERSEILUNG (SUB)	–	VERSEILEN (VRB)
ROTIERBAR (ADJ)	–	ROTIEREN (VRB)
		ROTATION (SUB)
DICHTEN (VRB)	–	DICHTUNG (SUB)

The automatic description generated 5,216 descriptors, that means about 326 descriptors per document. 3,959 simple descriptors (1,070 different ones) were generated out of the

3,832 words belonging to the word classes noun, adjective, and verb - including the decompositions.

Because of the identified (noun) structures another 1.259 precoordinations (so-called Complex descriptors) were generated. They are distributed as follows:

type	quantity	example
adj./noun	442	ISOLIERENDE ABSPERRMASSE
noun KON noun	202	DRAHT K BAND
noun GEN noun	236	HERSTELLUNG G LEITERBUENDEL
noun PRP noun	148	KORROSIONSSCHUTZ P KABELMANTEL
noun verb	158	DRAHT ERWAERMEN
verb MOD verb	71	VERARBEITEN KOENNEN

The CPU time used in total (on the computer TR 440 of the university of Saarland) was about 1,723 CPU seconds, that is 7.6 seconds/sentence or 0.26 seconds/word. The morphological analysis came to 433 seconds (25.1%), the generation of the descriptors was about 320 seconds (18.6%).

4. The connection between the analysis of corpora, the simulation of grammatical models, and the practical application

The practical application (even as a test) of the automatic indexing procedure which is presented here, and its integration into an information retrieval system has a long story. In the beginning of the sixties, research work concerning the syntax of the German language was started by Hans Eggers. At that time, the target was to develop a grammar of "Basic German", especially as a grammar for foreigners. The textual basis was then a corpus of 50,000 sentences by 50 authors, that is 1,000 sentences by each author, taken from the "Rowohlt's Deutsche Enzyklopädie" (RDE) /5/, respectively 50 authors of the "Frankfurter Allgemeine Zeitung" (FAZ). All sentences containing 4, 8, 16, 24, and 32 words were selected; their syntactical structure was described intellectually (according to the school grammar); ensuing, the results were evaluated by means of punched cards. By relating the text words to the coded word classes, an index of words and word classes was produced (among others).

In the middle of the sixties - using for the first time a computer of the University of Saarland - a simulation model guided along the surface was developed for an automatic description of the structure of German sentences. As a partial quantity (on the basis of the word inventory of the RDE and FAZ exploitation) an automatic dictionary of word forms was produced; the grammar was guided along surface valencies. This procedure showed clearly the capacities of the automatic syntactical analysis /6/; to some regards, however (coding of the dictionary, computer language, methods for the analysis of homographs etc.) it was unwieldy or unsatisfactory.

As a new start in the beginning of the seventies, a project was started in Saarbrücken within the scope of the "Sonderforschungsbereich Elektronische Sprachforschung" - mainly on the basis of findings of transformational grammar and by integrating parts of the valency theory. On the one

hand, the target was marked by a more systematic lexical approach (construction and coding of a morpho-syntactical dictionary containing about 80,000 items); on the other hand, the aim was the development of an automatic analysis system /7/, above all including the integration of automatic translation. In more than 10 years of research work the language analysis system was developed; it is now the basis of the CTX system. In the middle of the seventies the research work for the application of this system were started - especially for the German language analysis; the target was the use of the system in documentation (automatic indexing) /8/. After the beginnings in Regensburg, the project now continues within the scope of the field "information science" which was established in Saarbrücken in 1980.

Even when leaving out of consideration the roundabout ways and the dead ends which were followed during the basic research works in this area, the expenditure of work for the development of this system is about 120 - 150 "person years". This means capital expenditure of already more than 5 million DM. One has to consider this fact when working in this area.

In these days it appears in outlines that the system for automatic language analysis and indexing developed in Saarbrücken will find the "breakthrough" into the practical application for the retrieval of text contents. Probably the adaptations to users' problems will require another 5 years.
As the research work leaded to new findings requiring the further development of parts of the system, a new system is going to be developed at the same time. This new one will be based on the newest findings in software technology as well as on new safe results of linguistic research works. A completely new start, however (as, for instance, in the lexical field) will not be necessary: In will be possible to build up on existing procedures, and to develop them systematically.

Especially the application near to the practice raised new questions which (as in the field of semantic disambiguation) require new strategies. Thus there is an interrelation between basic research and practical application. In contrast to abstract and theoretical linguistic studies of "universal" kind however - which are fully legitimated anyway - the requirements of practical application will be placed into the foreground: The problem is to deal with the flood of informations, and consequently with the language barrier: - this is a possibility to close the gap of information crisis.

/1/ BMFT-Leistungsplan Fachinformation, Planperiode 1982 - 1984, Bonn 1982, p. 9.

/2/ BMFT-Leistungsplan ..., p. 34.

/3/ Gerhard Knorz: Die Darmstädter Projekte zur Automatischen Indexierung: WAI und AIR. In: Das Inforum 11 (1981), p. 38-49. (ISSN 0720-3950)

/4/ Harald Zimmermann: Das Projekt JUDO-DS - Juristische Dokumentanalyse im Bereich Datenschutz. In: Das Inforum 11 (1981), p. 50-65.

/5/ Hans Eggers: Zur Syntax der deutschen Sprache der Gegenwart. In: Studium Generale 15 (1962) p. 49-49.

/6/ Hans Eggers et al.: Elektronische Syntaxanalyse der deutschen Gegenwartssprache. Tübingen 1969.

/7/ SALEM. Ein Verfahren zur automatischen Lemmatisierung deutscher Texte. Ed.: Sonderforschungsbereich 100. Tübingen 1980.

/8/ Kurzbeschreibung des Systems CTX. Veröffentlichungen der Fachrichtung Informationswissenschaft. Saarbrücken 1982.