IV. Automatic Processing of Foreign Language Documents

G. Salton

Abstract

Experiments conducted over the last few years with the SMART document retrieval system have shown that fully automatic text processing methods using relatively simple linguistic tools are as effective for purposes of document indexing, classification, search, and retrieval as the more elaborate manual methods normally used in practice. Up to now, all experiments were carried out entirely with English language queries and documents.

The present study describes an extension of the SMART procedures to German language materials. A multi-lingual thesaurus is used for the analysis of documents and search requests, and tools are provided which make it possible to process English language documents against German queries, and vice versa. The methods are evaluated, and it is shown that the effectiveness of the mixed language processing is approximately equivalent to that of the standard process operating within a single language only.

1. Introduction

For some years, experiments have been under way to test the effectiveness of automatic language analysis and indexing methods in information retrieval. Specifically, document and query tests are processed fully automatically, and content identifiers are assigned using a variety of linguistic

tools, including word stem analysis, thesaurus look-up, phrase recognition, statistical term association, syntactic analysis, and so on. The resulting concept identifiers assigned to each document and search request are then matched, and the documents whose identifiers are sufficiently close to the queries are retrieved for the user's attention.

The automatic analysis methods can be made to operate in real-time — while the customer waits for an answer — by restricting the query-document comparisons to only certain document classes, and interactive user-controlled search methods can be implemented which adjust the search request during the search in such a way that more useful, and less useless, material is retrieved from the file.

The experimental evidence accumulated over the last few years indicates that retrieval systems based on automatic text processing methods — including fully automatic content analysis as well as automatic document classification and retrieval — are not in general inferior in retrieval effectiveness to conventional systems based on human indexing and human query formulation.

One of the major objections to the practical utilization of the automatic text processing methods has been the inability automatically to handle foreign language texts of the kind normally stored in documentation and library systems. Recent experiments performed with document abstracts and search requests in French and German appear to indicate that these objections may be groundless.

In the present study, the SMART document retrieval system is used to carry out experiments using as input foreign language documents and queries. The foreign language texts are automatically processed using a

thesaurus (synonym dictionary) translated directly from a previously available English version. Foreign language query and document texts are looked-up in the foreign language thesaurus and the analyzed forms of the queries and documents are then compared in the standard manner before retrieving the highly matching items. The language analysis methods incorporated into the SMART system are first briefly reviewed. Thereafter, the main procedures used to process the foreign language documents are described, and the retrieval effectiveness of the English text processing methods is compared with that of the foreign language material.

2. The SMART System

SMART is a fully-automatic document retrieval system operating on the IBM 7094 and 360 model 65. Unlike other computer-based retrieval systems, the SMART system does not rely on manually assigned key words or index terms for the identification of documents and search requests, nor does it use primarily the frequency of occurrence of certain words or phrases included in the texts of documents. Instead, an attempt is made to go beyond simple word-matching procedures by using a variety of intellectual aids in the form of synonym dictionaries, hierarchical arrangements of subject identifiers, statistical and syntactic phrase generation methods and the like, in order to obtain the content identifications useful for the retrieval process.

Stored documents and search requests are then processed without any prior manual analysis by one of several hundred automatic content analysis methods, and those documents which most nearly match a given search request are extracted from the document file in answer to the request. The system may be controlled by the user, in that a search request can be processed

first in a standard mode; the user can then analyze the output obtained and, depending on his further requirements, order a reprocessing of the request under new conditions. The new output can again be examined and the process iterated until the right kind and amount of information are retrieved. [1,2,3]

SMART is thus designed as an experimental automatic retrieval system of the kind that may become current in operational environments some years hence. The following facilities, incorporated into the SMART system for purposes of document analysis may be of principal interest:

a) a system for separating English words into stems and affixes (the so-called suffix 's' and stem thesaurus methods) which can be used to construct document identifications consisting of the stems of words contained in the documents;

b) a synonym dictionary, or thesaurus, which can be used to recognize synonyms by replacing each word stem by one or more "concept" numbers; these concept numbers then serve as content identifiers instead of the original word stems;

c) a hierarchical arrangement of the concepts included in the thesaurus which makes it possible, given any concept number, to find its "parents" in the hierarchy, its "sons", its "brothers", and any of a set of possible cross references; the hierarchy can be used to obtain more general content identifiers than the ones orignally given by going up in the hierarchy, more specific ones by going down, and a set of related ones by picking up brothers and cross-references;

d) statistical procedures to compute similarity coefficients based on co-occurrences of concepts within the sentences of a given collection; the related concepts, determined by statistical association, can then be added to the originally available concepts to identify the various documents;

e) syntactic analysis methods which make it possible to compare

the syntactically analyzed sentences of documents and search requests with a pre-coded dictionary of syntactic structures ("criterion trees") in such a way that the same concept number is assigned to a large number of semantically equivalent, but syntactically quite different constructions;

f) statistical phrase matching methods which operate like the preceding syntactic phrase procedures, that is, by using a preconstructed dictionary to identify phrases used as content identifiers; however, no syntactic analysis is performed in this case, and phrases are defined as equivalent if the concept numbers of all components match, regardless of the syntactic relationships between components;

g) a dictionary updating system, designed to revise the several dictionaries included in the system:

   i) word stem dictionary
   ii) word suffix dictionary
   iii) common word dictionary (for words to be deleted during analysis)
   iv) thesaurus (synonym dictionary)
   v) concept hierarchy
   vi) statistical phrase dictionary
   vii) syntactic ("criterion") phrase dictionary.

The operations of the system are built around a supervisory system which decodes the input instructions and arranges the processing sequence in accordance with the instructions received. The SMART system's organization makes it possible to evaluate the effectiveness of the various processing methods by comparing the outputs produced by a variety of different runs. This is achieved by processing the same search requests against the same document collections several times, and making judicious changes in the analysis procedures between runs. In each case, the search effectiveness is evaluated by presenting paired comparisons of the average performance over many search requests for two given search and retrieval methodologies.

3. The Evaluation of Language Analysis Methods

Many different criteria may suggest themselves for measuring the performance of an information system. In the evaluation work carried out with the SMART system, the effectiveness of an information system is assumed to depend on its ability to satisfy the users' information needs by retrieving wanted material, while rejecting unwanted items. Two measures have been widely used for this purpose, known as recall and precision, and representing respectively the proportion of relevant material actually retrieved, and the proportion of retrieved material actually relevant. [3] (Ideally, all relevant items should be retrieved, while at the same time, all nonrelevant items should be rejected, as reflected by perfect recall and precision values equal to 1).

It should be noted that both the recall and precision figures achievable by a given system are adjustable, in the sense that a relaxation of the search conditions often leads to high recall, while a tightening of the search criteria leads to high precision. Unhappily, experience has shown that on the average recall and precision tend to vary inversely since the retrieval of more relevant items normally also leads to the retrieval of more irrelevant ones. In practice, a compromise is usually made, and a performance level is chosen such that much of the relevant material is retrieved, while the number of nonrelevant items which are also retrieved is kept within tolerable limits.

In theory, one might expect that the performance of a retrieval system would improve as the language analysis methods used for document and query processing become more sophisticated. In actual fact, this turns out not to be the case. A first indication of the fact that retrieval effec-

tiveness does not vary directly with the complexity of the document or query analysis was provided by the output of the Aslib-Cranfield studies. This project tested a large variety of indexing languages in a retrieval environment, and came to the astonishing conclusion that the simplest type of indexing language would produce the best results. [4] Specifically, three types of indexing languages were tested, called respectively single terms, (that is, individual terms, or concepts assigned to documents and queries), controlled terms (that is, single terms assigned under the control of the well-known EJC Thesaurus of Engineering and Scientific Terms), and finally simple concepts (that is, phrases consisting of two or more single terms). The results of the Cranfield tests indicated that single terms are more effective for retrieval purposes than either controlled terms, or complete phrases. [4]

These results might be dismissed as being due to certain peculiar test conditions if it were not for the fact that the results obtained with the automatic SMART retrieval system substantially confirm the earlier Cranfield output. [3] Specifically, the following basic conclusions can be drawn from the main SMART experiments:

a) the simplest automatic language analysis procedure consisting of the assignment to queries and documents of weighted word stems originally contained in these documents, produces a retrieval effectiveness almost equivalent to that obtained by intellectual indexing carried out manually under controlled conditions; [3,5]

b) use of a thesaurus look-up process, designed to recognize synonyms and other term relations by replacing the original word stems by the corresponding thesaurus categories, improves the retrieval effectiveness by about ten percent in both recall and

precision;

c) additional, more sophisticated language analysis procedures, including the assignment of phrases instead of individual terms, the use of a concept hierarchy, the determination of syntactic relations between terms, and so on, do not, on the average, provide improvements over the standard thesaurus process.

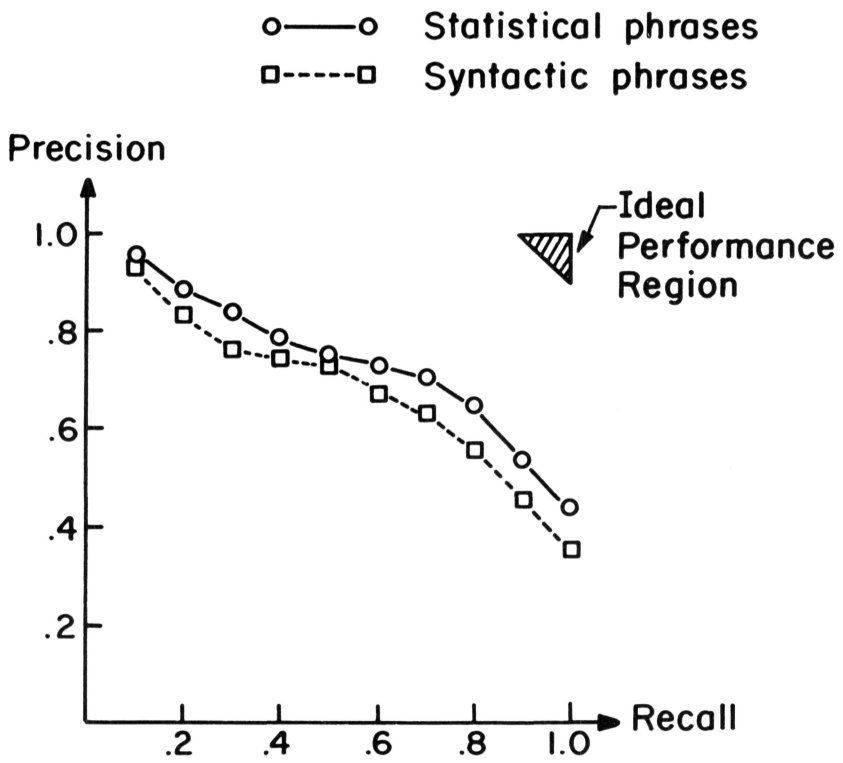
An example of a typical recall-precision graph produced by the SMART system is shown in Fig. 1, where a statistical phrase method is compared with a syntactic phrase procedure. In the former case, phrases are assigned as content identifiers to documents and queries whenever the individual phrase components are all present within a given document; in the latter case, the individual components must also exhibit an appropriate syntactic relationship before the phrase is assigned as an identifier. The output of Fig. 1 shows that the use of syntax degrades performance (the ideal performance region is in the upper right-hand corner of the graph where both the recall and the precision are close to 1). Several arguments may explain the output of Fig. 1:

a) the inadequacy of the syntactic analyzer used to generate syntactic phrases;

b) the fact that phrases are often appropriate content identifiers even when the phrase components are not syntactically related in a given context (e.g. the sentence "people who need information, require adequate retrieval services" is adequately identified by the phrase "information retrieval", even though the components are not related in the sentence);

c) the variability of the user population which makes it unwise to overspecify document content;

d) the ambiguity inherent in natural language texts which may work to advantage when attempting to satisfy the information



Recall	Precision ○——○	Precision □-----□
0.1	.960	.938
0.3	.834	.776
0.5	.769	.735
0.7	.706	.625
0.9	.546	.467

Comparison Between Statistical and Syntactic Phrases
(averages over 17 queries)

Fig. 1

needs of a heterogeneous user population with diverse information needs.

Most likely a combination of some of the above factors is responsible for the fact that relatively simple content analysis methods are generally preferable in a retrieval environment to more sophisticated methods. The foreign language processing to be described in the remainder of this study must be viewed in the light of the foregoing test results.

4. Multi-lingual Thesaurus

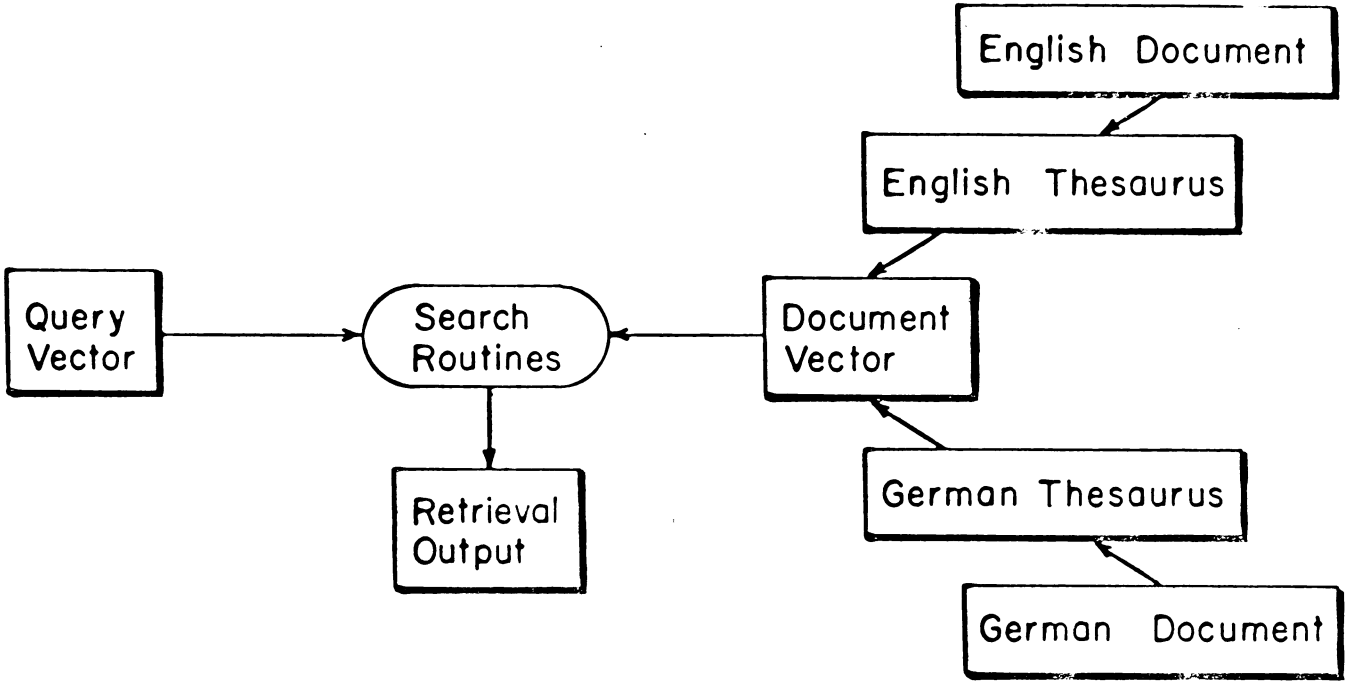
The multi-lingual text processing experiment is motivated by the following principal considerations:

a) in typical American libraries up to fifty percent of the stored materials may not be in English; about fifty percent of the material processed in a test at the National Library of Medicine in Washington was not in English (of this, German accounted for about 25%, French for 23%, Italian for 13%, Russian for 11%, Japanese for 6%, Spanish for 5%, and Polish for 5%); [6]

b) in certain statistical text processing experiments carried out with foreign language documents, the test results were about equally good for German as for English; [7]

c) simple text processing methods appear to work well for English, and there is no a priori reason why they should not work equally well for another language.

The basic multi-lingual system used for test purposes is outlined in Fig. 2. Document (or query) texts are looked-up in a thesaurus and reduced to "concept vector" form; query vectors and document vectors are then compared, and document vectors sufficiently similar to the query are withdrawn from the file. In order to insure that mixed language input is properly processed, the thesaurus must assign the same concept categories, no matter what the input language. The SMART system therefore utilizes a



Foreign Language Text Processing System

Fig. 2

multi-lingual thesaurus in which one concept category corresponds both to a family of English words, or word stems, as well as to their German translation.

A typical thesaurus excerpt is shown in Fig. 3, giving respectively concept numbers, English word class, and corresponding German word class. This thesaurus was produced by manually translating into German an originally available English version. Tables 1 and 2 show the results of the thesaurus look-up operation for the English and German versions of query QB 13. The original query texts in three languages (English, French, and German) are shown in Fig. 4. It may be seen that seven out of 9 "English" concepts are common with the German concept vector for the same query. In view of this, one may expect that the German query processed against the German thesaurus could be matched against English language documents as easily as the English version of the query. Tables 1 and 2 also show that more query words were not found during look-up in the German thesaurus than in the English one. This is due to the fact that only a preliminary incomplete version of the German thesaurus was available at run time.

5. Foreign Language Retrieval Experiment

To test the simple multi-lingual thesaurus process two collections of documents in the area of library science and documentation (the Ispra collection) were processed against a set of 48 search requests in documentation area. The English collection consisted of 1095 document abstracts, whereas the German collection contained only 468 document abstracts. The overlap between the two collections included 50 common documents. All 48 queries were originally available in English; they were manually translated

230	ART	ARCHITEKTUR
231	INDEPEND	SELBSTAENDIG UNABHAENGIG
232	ASSOCIATIVE	
233	DIVIDE	
234	ACTIVE	AKTIV
	ACTIVITY	AKTIVITAET
	USAGE	TAETIGKEIT
235	CATHODE	DIODE
	CRT	VERZWEIGER
	DIODE	
	FLYING-SPOT	
	RAY	
	RELAIS	
	RELAY	
	SCANNER	
	TUBE	
236	REDUNDANCY	
	REDUNDANT	
237	CHARGE	EINGANG
	ENTER	EINGEGANGEN
	ENTRY	EINGEGEBEN
	INSERT	EINSATZ
	POST	EINSTELLEN
		EINTRAGUNG
238	MULTI-LEVEL	
	MULTILEVEL	
239	INTELLECT	GEISTIG
	INTELLECTUAL	
	INTELLIG	
	MENTAL	
	MIND	
	NON-INTELLECTUAL	
240	ACTUAL	PRAXIS
	PRACTICE	
	REAL	

Excerpt from Multi-Lingual Thesaurus

Fig. 3

English Query QB 13

Concepts	Weights	Thesaurus Category
3 ✓	12	computer, processor
19 ✓	12	automatic, semiautomatic
33 ✓	12	analyze, analyzer, analysis, etc.
49	12	compendium, compile, deposit
65 ✓	12	authorship, originator
147 ✓	12	discourse, language, linguistic
207 ✓	12	area, branch, subfield
267 ✓	12	concordance, keyword-in-context, KWIC
345	12	bell
*		anonymous, lettres

✓ common concept with German query

* words not found in thesaurus

Thesaurus Look-up for English Query QB 13

Table 1

German Query QB 13

Concepts	Weights	Thesaurus Category
3 ✓	12	Computer, Datenverarbeitung
19 ✓	12	Automatisch, Kybernetik
21	4	Artikel, Presse, Zeitschrift
33 ✓	6	Analyse, Sprachenanalyse
45	4	Herausgabe, Publikation
64	4	Buch, Heft, Werk
65 ✓	12	Autor, Verfasser
68	12	Literatur
147 ✓	6	Linguistik, Sprache
207 ✓	12	Arbeitsgebiet, Fach
267 ✓	12	Konkordanz, KWIC
*		schoenen, hilfreich, vermutlich anonymen, zusammenzustellen

✓ common concept with English query

* words not found in thesaurus

Thesaurus Look-up for German Query QB 13

Table 2

*FIND Q13BAUTHORS

IN WHAT WAYS ARE COMPUTER SYSTEMS BEING APPLIED TO RESEARCH IN THE FIELD OF THE BELLES LETTRES ? HAS MACHINE ANALYSIS OF LANGUAGE PROVED USEFUL FOR INSTANCE, IN DETERMINING PROBABLE AUTHORSHIP OF ANONYMOUS WORKS OR IN COMPILING CONCORDANCES ?

DANS QUEL SENS LES CALCULATEURS SONT-ILS APPLIQUES A LA RECHERCHE DANS LE DOMAINE DES BELLES-LETTRES ? EST-CE QUE L'ANALYSE AUTOMATIQUE DES TEXTES A ETE UTILE, PAR EXEMPLE, POUR DETERMINER L'AUTEUR PROBABLE D'OUVRAGES ANONYMES OU POUR FAIRE DES CONCORDANCES ?

INWIEWEIT WERDEN COMPUTER-SYSTEME ZUR FORSCHUNG AUF DEM GEBIET DER SCHUENEN LITERATUR VERWENDET ? HAT SICH MASCHINELLE SPRACHENANALYSE ALS HILFREICH ERWIESEN, UM Z.B. DIE VERMUTLICHE AUTORENSCHAFT BEI ANONYMEN WERKEN ZU BESTIMMEN ODER UM KONKORDANZEN ZUSAMMENZUSTELLEN ?

Query QB 13 in Three Languages

Fig. 4

into German by a native German speaker. The English queries were then processed against both the English and the German collections (runs E-E and E-G), and the same was done for the translated German queries (runs G-E and G-G, respectively). Relevance assessments were made for each English document abstract with respect to each English query by a set of eight American students in library science, and the assessors were not identical to the users who originally submitted the search requests. The German relevance assessments (German documents against German queries), on the other hand, were obtained from a different, German speaking, assessor.

The principal evaluation results for the four runs using the thesaurus process are shown in Fig. 5, averaged over 48 queries in each case. It is clear from the output of Fig. 5 that the cross-language runs, E-G (English queries - German documents) and G-E (German queries - English documents), are not substantially inferior to the corresponding output within a single language (G-G and E-E, respectively), the difference being of the order of 0.02 to 0.03 for a given recall level. On the other hand, both runs using the German document collection are inferior to the runs with the English collection.

The output of Fig. 5 leads to the following principal conclusions:

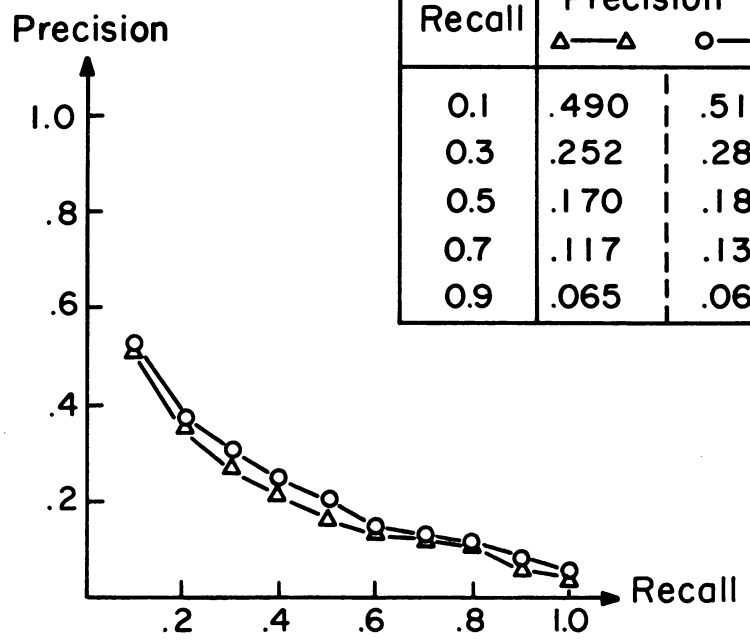
a) the query processing is comparable in both languages; for if this were not the case, then one would expect one set of query runs to be much less effective than the other (that is, either E-E and E-G, or else G-G and G-E);

b) the language processing methods (that is, thesaurus categories, suffix cut-off procedures, etc.) are equally effective in both cases; if this were not the case, one would expect one of the single language runs to come out very poorly, but

△——△ English queries
German documents

o——o German queries
German documents

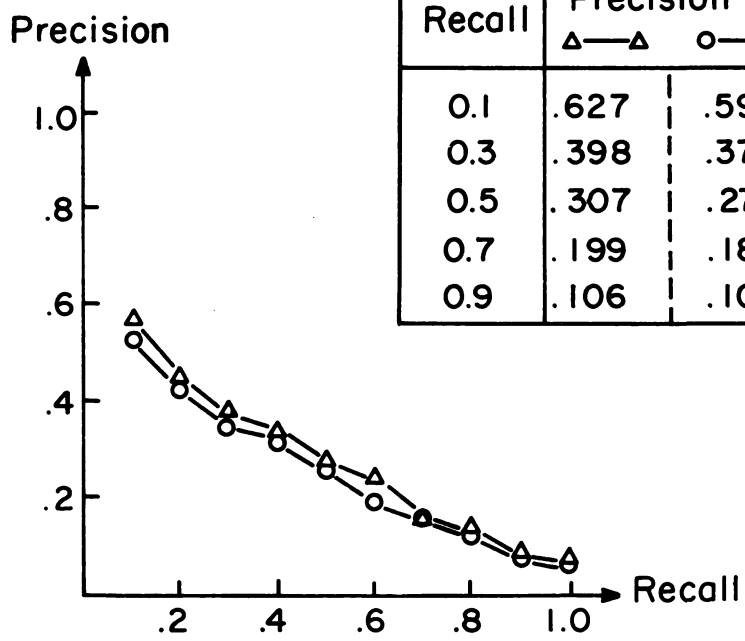
Recall	Precision △——△	Precision o——o
0.1	.490	.513
0.3	.252	.286
0.5	.170	.181
0.7	.117	.130
0.9	.065	.066



△——△ English queries
English documents

o——o German queries
English documents

Recall	Precision △——△	Precision o——o
0.1	.627	.599
0.3	.398	.374
0.5	.307	.276
0.7	.199	.184
0.9	.106	.100



Basic Comparison English vs. German Queries (Thesaurus Process)

Fig. 5

neither E-E, nor G-G came out as the poorest run;

c) the cross-language runs are performed properly, for if this were not the case, one would expect E-G and G-E to perform much less well than the runs within a single language; since this is not the case, the principal conclusion is then obvious that documents in one language can be matched against queries in another nearly as well as documents and queries in a single language;

d) the runs using the German document collection (E-G and G-G) are less effective than those performed with the English collection; the indication is then apparent that some characteristic connected with the German document collection itself — for example, the type of abstract, or the language of the abstract, or the relevance assessments — requires improvement; the effectiveness of the cross-language processing, however, is not at issue.

The foreign language analysis is summarized in Table 3.

6. Failure Analysis

Since the query processing operates equally well in both languages, while the German document collection produces a degraded performance, it becomes worthwhile to examine the principal differences between the two document collections. These are summarized in Table 4. The following principal distinctions arise:

a) the organization of the thesaurus used to group words or word stems into thesaurus categories;

b) the completeness of the thesaurus in terms of words included in it;

c) the type of document abstracts included in the collection;

Translation Problem	Corresponding Observation	Observation Confirmed
Poor query processing or poor translation	E-E and E-G much better than G-E and G-G, or vice versa	No
Poor language processing	Either E-E or G-G much poorer than cross-language runs	No
Poor cross-language processing	Both E-G and G-E poorer than other runs	No
Poor processing of one document collection	Either E-G and G-G, or else G-E and E-E simultaneously poor	Yes

E-E: English queries - English documents
E-G: English queries - German documents
G-E: German queries - English documents
G-G: German queries - German documents

Analysis of Foreign Language Processing

Table 3

Characteristic of Collections	Document Collection	
	English	German
Number of document abstracts	1095	468
Number of documents common to both collections	50	50
Number of queries used in test	48	48
Number of relevance assessors	8	1
Number of common relevance assessors	0	0
Generality of collection (number of relevant documents over total number of documents in collection)	0.013	0.029
Average number of word occurrences not found in the thesaurus during look-up of document abstracts	6.5	15.5

Characteristics of Document Collections

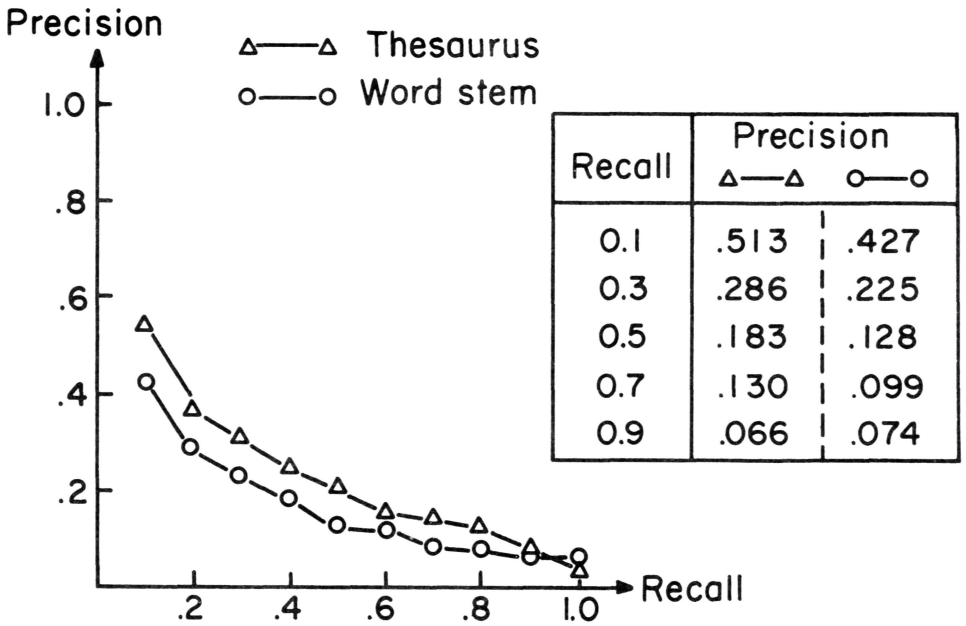
Table 4

d) the accuracy of the relevance assessments obtained from the collections.

Concerning first the organization of the multi-lingual thesaurus, it does not appear that any essential difficulties arise on that account. This is confirmed by the fact that the cross-language runs operate satisfactorily, and by the output of Fig. 6(a) comparing a German word stem run (using standard suffix cut-off and weighting procedures) with a German thesaurus run. It is seen that the German thesaurus improves performance over word stems for the German collection in the same way as the English thesaurus was seen earlier to improve retrieval effectiveness over the English word stem analysis. [2,3]

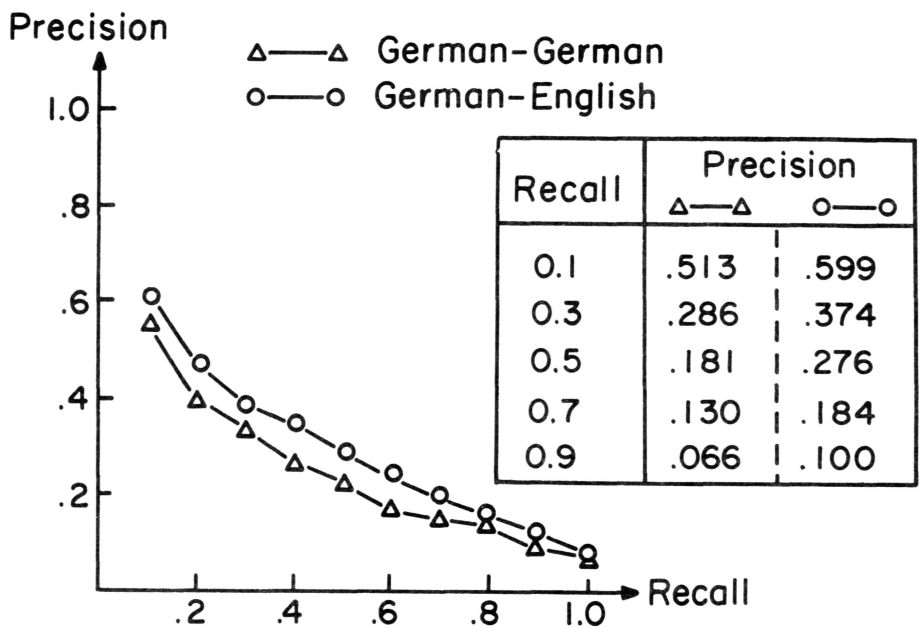
The other thesaurus characteristic — that is its completeness— appears to present a more serious problem. Table 4 shows that only approximately 6.5 English words per document abstract were not included in the English thesaurus, whereas over 15 words per abstract were missing from the German thesaurus. Obviously, if the missing words turn out to be important for content analysis purposes, the German abstracts will be more difficult to analyze than their English counterpart. A brief analysis confirms that many of the missing German words, which do not therefore produce concept numbers assignable to the documents, are indeed important for content identification. Fig. 7, listing the words not found for document 005, shows that 12 out of 14 missing words appear to be important for the analysis of that document. It would therefore seem essential that a more complete thesaurus be used under operational conditions and for future experiments.*

*A rerun of the experiment using a more complete thesaurus is decribed in the appendix.



Recall	Precision △—△	Precision o—o
0.1	.513	.427
0.3	.286	.225
0.5	.183	.128
0.7	.130	.099
0.9	.066	.074

a) Thesaurus Test for German Queries vs. German Documents



Recall	Precision △—△	Precision o—o
0.1	.513	.599
0.3	.286	.374
0.5	.181	.276
0.7	.130	.184
0.9	.066	.100

b) Relevance Judgment Test for German Queries (Thesaurus Process)

Thesaurus and Relevance Judgment Test

Fig. 6

68 PAGE 483 NOVEMBER 6, 1968 PAGE 484 00001

*TEXT 00501063 ZUR KOMPILATION VON THESAURI

WORD NOT FOUND	KIND	LOC	NUM	SENTENCE AND WORD NUMBERS
KOMPILATION	SUFFIX	2	1	1, 2
THESAURUSELEMENTE	SUFFIX	4	1	2, 5
GEBRAUCHT	SUFFIX	1	1	2, 13
SCHLAGWORTES	SUFFIX	2	1	3, 11
HOMONYME	SUFFIX	2	1	4, 4
SYNONYME	SUFFIX	2	1	4, 6
VERVOLLSTAENDIGUNG	SUFFIX	4	1	5, 3
VERWEISSYSTEMS	SUFFIX	4	1	5, 8
VORORDNUNG	SUFFIX	2	1	5, 13
HAUPTLISTE	SUFFIX	2	1	5, 15
HILFSLISTEN	SUFFIX	2	1	5, 19
UNERLAESSLICH	SUFFIX	6	1	5, 20
GEBRAUCH	SUFFIX	1	1	6, 11
KONKORDANZEN	SUFFIX	2	1	6, 15

List of Words not Found in Thesaurus
for Document 005

Fig. 7

The other two collection characteristics, including the type of abstracts and the accuracy of the relevance judgments are more difficult to assess, since these are not subject to statistical analysis. It is a fact that for some of the German documents informative abstracts are not available. For example, the abstract for document 028, included in Fig. 8, indicates that the corresponding document is a conference proceedings; very little is known about the subject matter of the conference, but the document was nevertheless judged relevant to six different queries (nos. 17, 27, 31, 32, 52, and 53) dealing with subjects as diverse as "behavioral studies of information system users" (query 17), and "the study of machine translation" (query 27). One might quarrel with such relevance assessments and with the inclusion of such documents in a test collection, particularly also since Fig. 6(b) shows that the German queries operate more effectively with the English collection (using English relevance assessments) than with the German assessments. However, earlier studies using a variety of relevance assessments with the same document collection have shown that recall-precision results are not affected by ordinary differences in relevance assessments. [8] For this reason, it would be premature to assume that the performance differences are primarily due to distinctions in the relevance assessments or in the collection make-up.

7. Conclusion

An experiment using a multi-lingual thesaurus in conjunction with two different document collections, in German and English respectively, has shown that cross-language processing (for example, German queries against English documents) is nearly as effective as processing within a single language. Furthermore, a simple translation of thesaurus categories appears

```
*TEXT 02401410 ANFORDERUNGEN DER PRAXIS AN DIE BEGRIFFSORDNUNG EINER
$ANFORDERUNGEN DER PRAXIS AN DIE BEGRIFFSORDNUNG EINER
$FACHDOKUMENTATION
ANFORDERUNGEN DER PRAXIS AN DIE BEGRIFFSORDNUNG EINER FACHDOKUMENTATION
. AUS DER PRAKTISCHEN ARBEIT FUER DEN AUFBAU EINER DOKUMENTATIONSSTELLE
UND IHREN TAEGLICHEN ANFORDERUNGEN HAT DER AUTOR DIE RICHTLINIEN EINER
BEGRIFFLICHEN ORDNUNG ENTWICKELT . SACHLICHE UND FINANZIELLE REALITAETEN
MUESSEN IN JEDEM EINZELFALL DEN IDEELLEN FORDERUNGEN GEGENUEBERGESTELLT
WERDEN . DREI FORDERUNGEN HAELT DER VERFASSER FUER ENTSCHEIDEND .. 1 .
ORDNUNG OHNE ZEITAUFWAND, 2 . UEBERSICHTLICHKEIT UND 3 . ELASTIZITAET
DES SYSTEMS UND SEINER ANWENDUNG . DIE BEGRIFFSORDNUNG DES PRAKTIKERS
ENTSTEHT AUS DEM BEDARF, NICHT AUS DER THEORIE .
```

```
*TEXT 02801548 INTERNATIONALER KONGRESS UND AUSSTELLUNG UEBER WISSENSCHA
$INTERNATIONALER KONGRESS UND AUSSTELLUNG UEBER WISSENSCHAFTLICHE
$UND TECHNISCHE DOKUMENTATION UND INFORMATION IN ROM VOM 2 . BIS
$11 . FEBRUAR 1964
INTERNATIONALER KONGRESS UND AUSSTELLUNG UEBER WISSENSCHAFTLICHE UND
TECHNISCHE DOKUMENTATION UND INFORMATION IN ROM VOM 2 . BIS 11 . FEBRUAR
1964 . DER KONGRESS WURDE VON DER ITALIENISCHEN PRODUKTIVITAETSZENTRALE
IN ROM DURCHGEFUEHRT . IN DIESEM ZUSAMMENHANGE WURDE DIE BUNDESREPUBLIK
DEUTSCHLAND DURCH DAS RKW IN SEINER EIGENSCHAFT ALS DEUTSCHE PZ
VERTRETEN . DER AUTOR GIBT EINE UEBERSICHT UEBER DIE HAUPTVORTRAEGE UND
DIE RAHMENVERANSTALTUNGEN DES KONGRESSES . ER HEBT BESONDERS DIE
TEILNAHME HOHER REGIERUNGSSTELLEN HERVOR .
```

Typical German Document Texts

Fig. 8

to produce a document content analysis which is equally effective in English as in German. In particular, differences in morphology (for example, in the suffix cut-off rules), and in language ambiguities do not seem to cause a substantial degradation when moving from one language to another. For these reasons, the automatic retrieval methods used in the SMART system for English appear to be applicable also to foreign language material.

Future experiments with foreign language documents should be carried out using a thesaurus that is reasonably complete in all languages, and with identical query and document collections for which the same relevance judgments may then be applicable across all runs.

References

[1] G. Salton and M. E. Lesk, The SMART Automatic Document Retrieval System — An Illustration, Communications of the ACM, Vol. 8, No. 6, June 1965.

[2] G. Salton, Automatic Information Organization and Retrieval, McGraw Hill Book Company, New York, 1968, 514 pages.

[3] G. Salton and M. E. Lesk, Computer Evaluation of Indexing and Text Processing, Journal of the ACM, Vol. 15, No. 1, January 1968.

[4] C. W. Cleverdon and E. M. Keen, Factors Determining the Performance of Indexing Systems, Vol. 1: Design, Vol. 2: Test Results, Aslib-Cranfield Research Project, Cranfield, England, 1966.

[5] G. Salton, A Comparison Between Manual and Automatic Indexing Methods, American Documentation, Vol. 20, No. 1, January 1969.

[6] F. W. Lancaster, Evaluation of the Operating Efficiency of Medlars, Final Report, National Library of Medicine, Washington, January 1969.

[7] J. H. Williams, Computer Classification of Documents, FID-IFIP Conference on Mechanized Documentation, Rome, June 1967.

[8] M. E. Lesk and G. Salton, Relevance Assessments and Retrieval System Evaluation, Information Storage and Retrieval, Vol. 4 No. 4, October 1968.

Appendix

To test the effect of the missing words in the German thesaurus, the experiments were repeated using a more complete thesaurus to which previoulsy missing entries had been added.

The following table summarizes the differences in results, averaged over 48 queries as before (see Fig. 5).

German Queries German Documents		
Recall	Precision	
	Old Thesaurus	New Thesaurus
.1	.513	.527
.3	.286	.327
.5	.181	.203
.7	.130	.140
.9	.066	.096

Average Precision at Fixed Recall Points

English Queries German Documents		
Recall	Precision	
	Old Thesaurus	New Thesaurus
.1	.490	.513
.3	.252	.299
.5	.170	.185
.7	.117	.122
.9	.065	.091

Average Precision at Fixed Recall Points

It may be noted that an improvement in average precision of 2 to 5 percent results from the dictionary change. Even after the dictionary replacement, the English collection produces better results than the German, the differences in precision having about 10 percent at most recall points. These differences are due to one or more of the following deficiencies:

a) unavailability of informative abstracts in German;

b) misspellings in the German text;

c) a program limitation which limits all German words to a limit of 24 characters (no English words exceed this limit);

d) discrepancies in the relevance judgments pertaining to the German collection;

e) suffixing problems in German, particularly those dealing with single letter suffixes such as 's', 'n', and 't'.

These problems may eventually be solved in further work with the German texts. The basic results pertaining to the cross-language processing are in any case unaffected.