V. Syntax in Text Analysis

S. F. Weiss

Abstract

This paper discusses various methods for the determination of phrases from text and for their use in information retrieval. The results of a set of experiments using these methods are presented and analyzed. Future work in this area is also covered.

1. Introduction

The purpose of this project is to investigate the use of syntax in in text analysis and information retrieval and specifically, to determine the usefulness of phrases in the retrieval process. In its present state the SMART system assigns concept vectors to documents partly on the basis of the words in the document. Two similar yet nonidentical documents can thus be assigned identical vectors, as in the case of the two sentences:

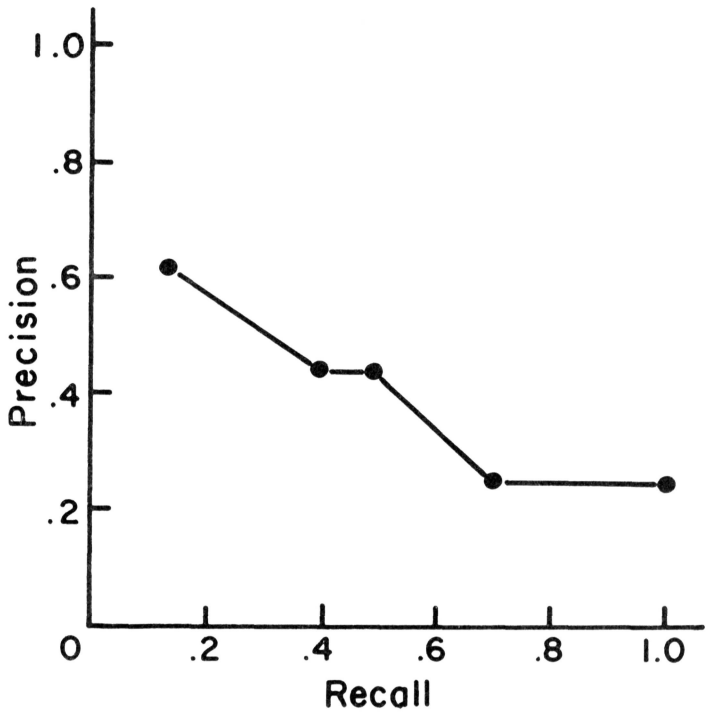
the programming of numerical analysis systems;
numerical analysis in programming systems.

If it is desired to differentiate between these two sentences, the obvious solution is to use syntax. What is not obvious is how this process is to be accomplished. Various possibilities exist; experiments with some of these methods are discussed in the following sections. The experiments are performed using the ADI collection of 82 documents. A set of ten queries, five general and five specific, is chosen as representative of the various forms and constructions of queries. A normal SMART run is

performed using the cosine coefficient. For each of the ten test queries, the ten most highly correlated documents are identified. These documents, along with any others, relevant to the test queries but not in the top ten, are collected to form the test document set. The total set contains 56 of the 82 ADI documents. In all the experiments, phrases are determined for this test set only. It is felt that the results achieved with the limited set will differ little from those of the full set. Also because of the great quantity of hand work required, the restricted document set is necessary. The results of the normal cosine retrieval process are shown in Fig. 1 in the form of the average precision-recall curve for the ten test queries.

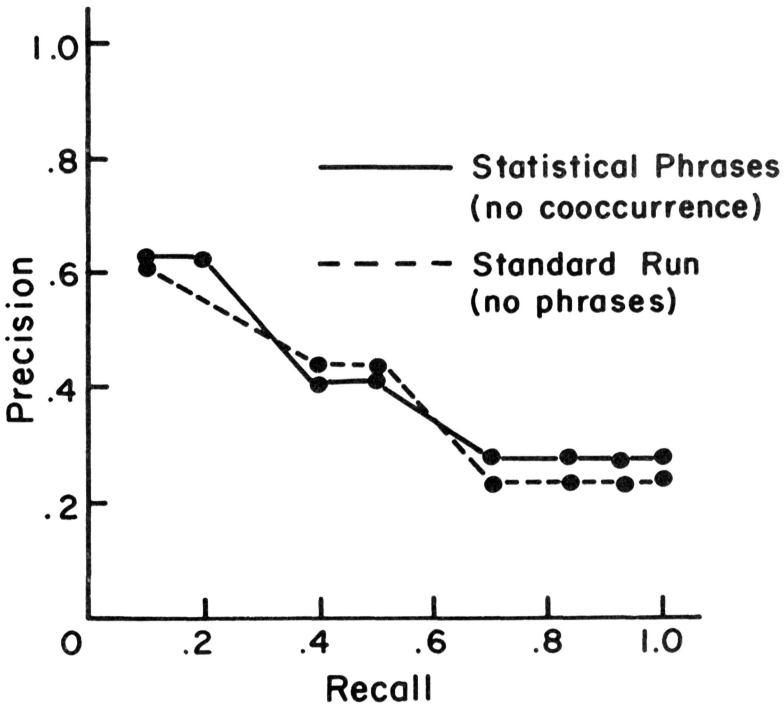
2. Statistical Phrases

The statistical phrase process uses a predetermined list of phrases. The cooccurrence of the phrase elements in a document is considered an occurrence of that phrase regardless of the syntactic relation of the phrase components. A concept number is associated with each phrase and the appropriate concepts appended to the document or query vectors. This method is clearly the simplest way to determine phrases since it requires no syntactic analysis of the text. It however has some rather serious drawbacks. Most obvious is the fact that it may recognize false phrases; that is occurrences of the desired phrase elements but not in the proper syntactic relation. This problem can be minimized in small collections dealing with a narrow subject area by judicious selection of the statistical phrase list. In a corpus dealing with computer systems, for example, the occurrence of the words "real" and "time" can be viewed with relative certainty to be an occurrence of the phrase "real time". However as the collection grows



Average Precision-Recall Curve for Experiment 1
(no phrases)

Fig. 1



Precision-Recall Curve for Experiment 2

Fig. 2

and the subject area broadens, these decisions become less certain. Also the difficulty in creating the phrase list is increased as the corpus is enlarged. The phrase list can be determined by statistical means, however weaknesses in this method can further compound the problem. In the ADI collection for example, of the 409 statistical phrases in the test document set, only 153, roughly 37%, are syntactically correct.

A second difficulty in using statistical phrases involves unmatched concept numbers. In many cases, especially when the document is long and the query is short, the document contains many more phrases than the query. Consequently many of the phrase concepts in the document are not matched with query concepts. These unmatched concepts clearly lower the correlation and partially if not completely, offset any gain achieved by the matched phrase concepts. Thus the inclusion of too many phrases dilutes the original vector with unusable information and inferior retrieval results are produced.

Fig. 2 shows a comparison of the average precison-recall values for retrieval with statistical phrases with those made with no phrases. The phrases are determined by hand searching the set of test documents and queries using the SMART list of statistical phrases. The weight assigned to a phrase concept is the minimum weight of its component concepts.

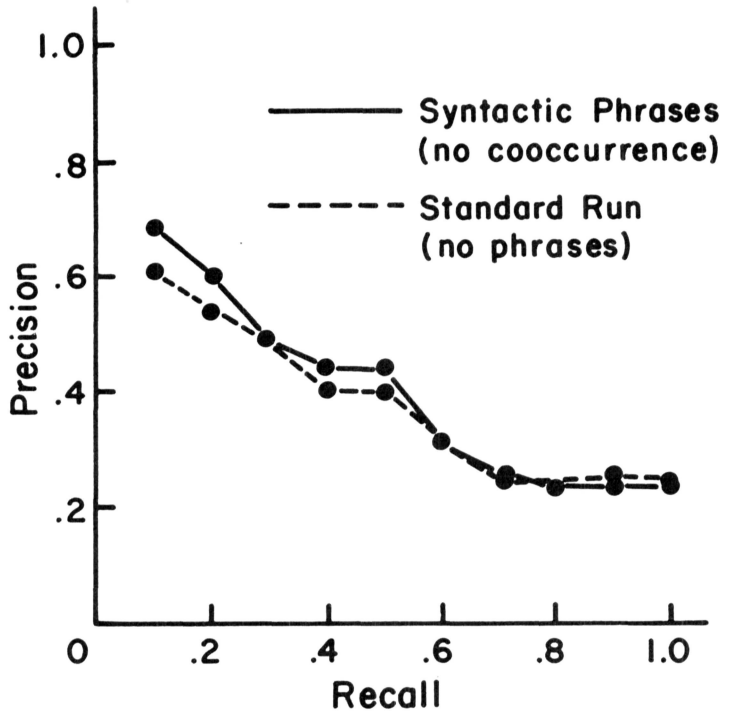
3. Syntactic Phrases

The text of the test documents and queries is searched again and those statistical phrases which are not syntactically correct are removed. This leaves 153 of the original 409 document phrases and 6 of the 12 query phrases. Results of retrieval using syntactic phrases are shown in Fig. 3

along with the results of the normal run. While the elimination of all syntactically false phrases solves one problem, another problem, the adverse effect of an overabundance of unmatched concepts, still remains. The following section presents a method for solution of this problem.

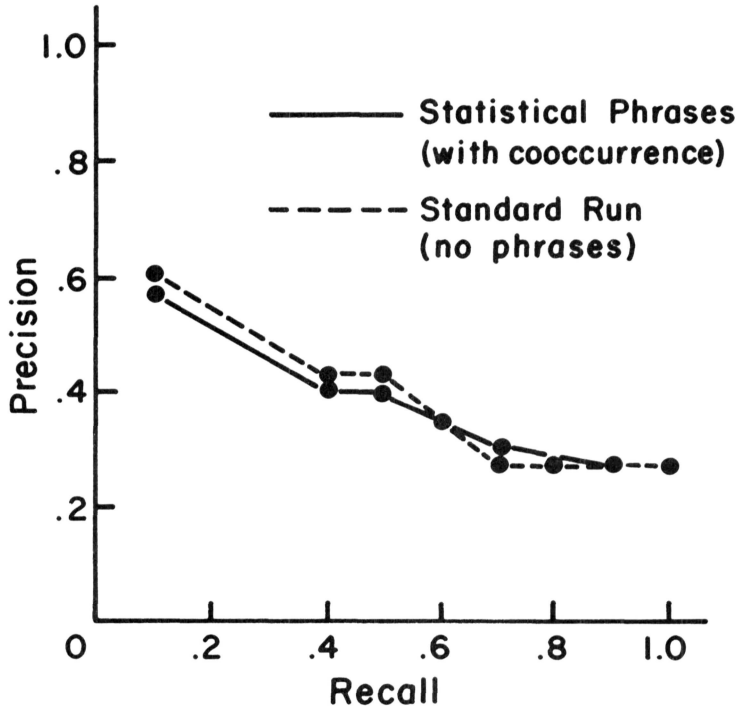
4. Cooccurrence

The methods described in sections 2 and 3 have a number of common problems. One is the dilution effect causes by unmatched concepts. A second difficulty deals with the value of a phrase as a nonrelevancy indicator. Individual concepts are about equally good as relevancy and nonrelevancy indicators. That is the cooccurrence of concept A in document D and query Q is as good a measure of D's relevance to Q as the lack of this co-occurrence is a measure of D's nonrelevance. As more and more structure is imposed on the comparison of documents and queries, high correlations grow more significant but less frequent while low correlations grow more frequent but less significant. For example if documents are retrieved only if they match, word for word, the complete query, few if any documents would ever be retrieved. However any document retrieved by this scheme would almost certainly be relevant. Thus the added structure makes the retrieved documents more likely to be relevant. On the other hand, the fact that some documents do not match the complete query is not a good indicator of their nonrelevancy. The situation is similar for phrases. The cooccurrence of a phrase in a document and query is a better indicator of the document's relevance to a query than two other cooccurring concepts that do not form a phrase. Furthermore an unmatched phrase concept is not as good a nonrelevancy measure as is an unmatched word concept. It is therefore necessary



Average Precision-Recall Curve for Experiment 3

Fig. 3



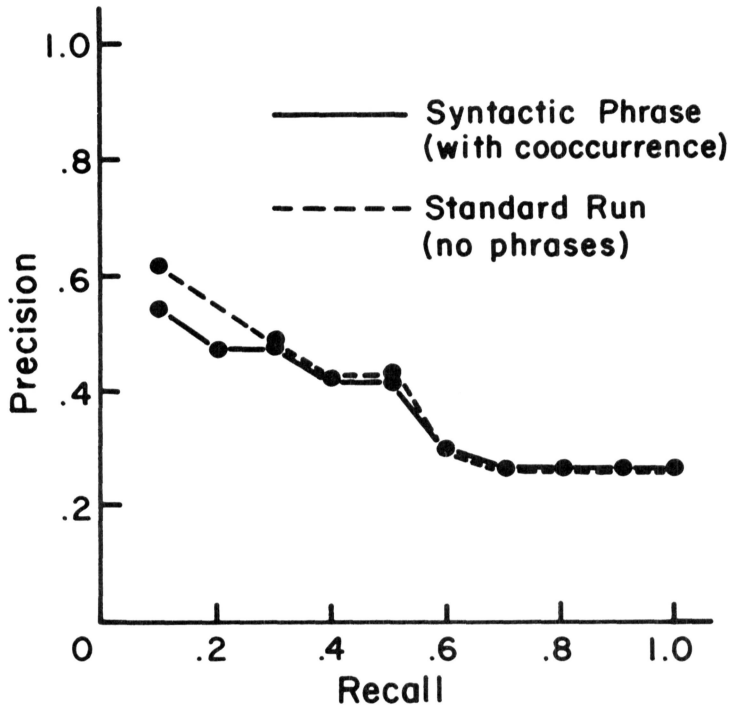
Average Precision-Recall Curve for Experiment 4

Fig. 4

to treat phrase concepts and word concepts differently.

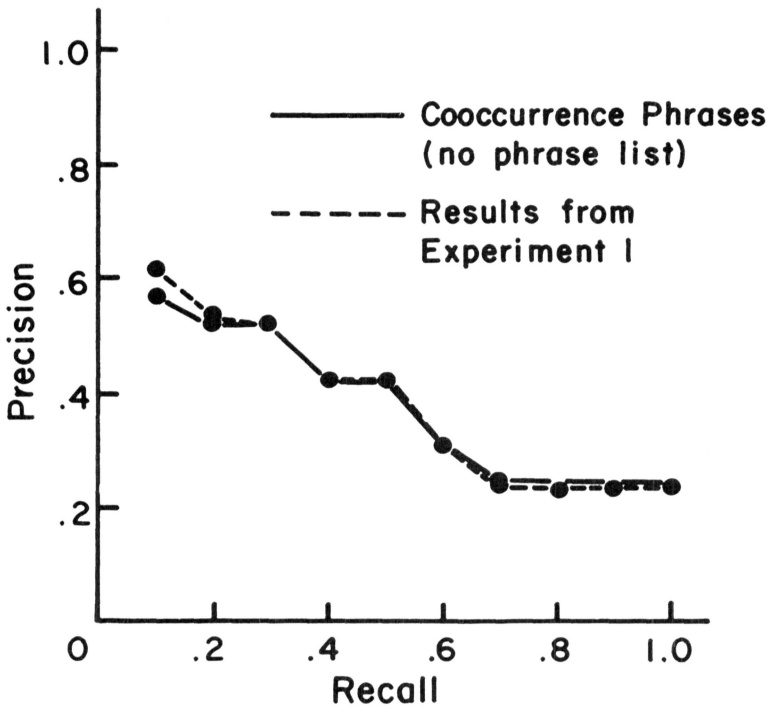
The method proposed to solve these difficulties is to consider phrases only when they cooccur in both the document and query. This is accomplished by the following procedure. Let D and Q be the word concept vectors for a given document and query, and PD and PQ, their associated phrase concept vectors. In the previous experiments the correlation is calculated between D + PD and Q + PQ. For the proposed process, the cooccurrence procedure first determines the set C = PD $\cap$ PQ. That is C is the set of all phrase concepts which occur in both the query and document. Correlation is then calculated using D + C and Q + C only. In this way it is guaranteed that phrase concepts cannot lower the correlation; and in the worst case when no phrases cooccur, C = $\emptyset$ and the correlation is the same as when no phrases are used. This process alleviates two of the major problems associated with phrase use. First, by ignoring all unmatched concepts, the vectors cannot become diluted and thus phrases cannot harm the retrieval. Secondly, phrases are used only as a relevancy indicator and their far weaker role of nonrelevancy indicator is ignored.

The next two experiments involve the use of statistical and syntactic phrases as previously determined, but incorporating the cooccurrence scheme. The correlation process for each document query pair involves forming a master document containing all the word concepts in the original vector as well as all phrase concepts common to both the document and query. A master query is similarly formed. The cosine coefficient is then calculated for the master document and master query. The results are shown in Fig. 4 and Fig. 5.



Average Precision-Recall Curve for Experiment 5

Fig. 5



Average Precision-Recall Curve for Experiment 6

Fig. 6

5. Elimination of the Phrase List

All methods discussed so far for using phrases in retrieval have required a phrase list. As previously mentioned the creation of these lists, whether by hand or by statistical processes, raises certain inherent problems. In general, it is far more desirable to be able to determine phrases without the need of such a list. One possible solution is to perform a syntactic analysis of the text, and determine all the phrases. The set of phrases thus generated is then normalized to associate all syntactically different but semantically identical phrases. This is accomplished, for example, by the use of a criterion tree-matching scheme. Each phrase in the reduced set is then assigned a concept number, and retrieval proceeds as in the previous cases. However the syntactic analysis and normalization processes are prohibitively complex and produce a very large number of phrases. For these reasons an alternate method is used.

The scheme used centers on the relation between pairs of concepts, and on locating cooccurrence phrases. A cooccurrence phrase is defined as any pair of concepts which occurs in both the document and the query and which has the same or equivalent syntactic relation in both. The text is first syntactically analyzed by hand. For each word in the text, the syntactic relation between it and every other word in the text is then determined. For a sentence of length n this can lead to a maximum of n(n-1) relations. In practice, however, the number of relations is far smaller than this because many words are not related. The types of relations considered are listed in Table 1.

After the relations are determined, each word is replaced by its concept number, and the two concepts and their relations are encoded

Code Number	Type of Relation
01	Modification
02	Parallel
03	Subject-Verb
04	Subject-Predicate nominative
05	Subject-Direct object
06	Verb-Object

The following rules are used to determine relations:

1. Modification
   a. In a string of adjectives followed by a noun, each adjective modifies all following adjectives as well as the noun.
   b. In the construction;

      ADJ NOUN-1 and NOUN-2

      the adjective modified both nouns. However, the case of

      NOUN-1 and ADJ NOUN-2

      the adjective modifies only the second noun.
   c. In the construction

      ADJ-1 and ADJ-2 NOUN

      both adjectives modify the noun.
   d. For certain prepositional phrases of the form

      NOUN-1 PREP NOUN PHRASE

      the entire noun phrase is treated as if it occurred in front of, and modifying NOUN-1. The most common preposition used in this way is "of" although others are also used.
   e. A predicate adjective modifies its subject

2. Parallel
   a. Two words of similar type and acting in similar roles are considered parallel. For example;

      Information storage and retrieval

      Fast and accurate method

      Neither programmers, analysts, nor researchers

3. All the other relations are fairly straightforward.

Relations Between Concepts

Table 1

XXXYYYZZ, where XXX is the first concept, YYY, the second, and ZZ, the relation between them. The order of the two concepts is significant for all relations except parallel; and for this case the smaller concept always appears first. The encoded concepts and relations are then treated as concept numbers and the retrieval process is carried out using the cooccurrence scheme. The results are seen in Fig. 6.

A system that uses a phrase list can recognize equivalent phrases whose constituent concepts are not equivalent. For example the phrases "memory holding" and "data processing" are both assigned the same phrase concept number by the SMART phrase list, while each of the four words falls into a different concept class. The recognition of such equivalent phrases is impossible for systems which do not employ such a list. It is therefore expected that the results of this experiment should be somewhat inferior to those achieved in the previous experiments. However, retrieval without the requirement of a phrase list seems to be a more reasonable approach to the problem especially in the case of large document collections.

6. Analysis of Results

The results of the six retrieval experiments are summarized in Fig. 7. The plus or minus to the right of each figure indicates whether it is above (+) or below (-) the value achieved for that recall level when no phrases are used (experiment 1). The results clearly show that there is no great gain achieved by the use of phrases and in some cases their use appears to be actually detrimental. However, upon more careful analysis of these results, a number of unusual factors are found which make these results less discouraging than they initially appear.

Recall	1	2	3	4	5	6
.1	.61242	.62538+	.64999+	.58761-	.56242-	.54575-
.2	.55242	.62538+	.59999+	.52761-	.50242-	.48575-
.3	.48618	.49573+	.47975-	.46385-	.48261-	.48618+-
.4	.42868	.40780-	.44225+	.42228-	.42437-	.42794-
.5	.43344	.40587-	.44701+	.40955-	.43270-	.43270-
.6	.32924	.33295+	.33162+	.32076-	.33376+	.33411+
.7	.25489	.28381+	.25514+	.26077+	.26165+	.25515+
.8	.25368	28781+	.24309-	.25543+	.25711+	.25061-
.9	.24133	27709+	.23509-	.24329+	.24914+	.24313+
1.0	.24133	27709+	.23509-	.24329+	.24914+	.24313+

Experiments:

1. No phrases
2. Statistical phrases, no cooccurrence
3. Syntactic phrases, no cooccurrence
4. Statistical phrases with cooccurrence
5. Syntactic phrases with cooccurrence
6. Cooccurrence phrases, no phrase lists

Precision Values at each Recall Level
for Each Experiment

Fig. 7

Consider first the results obtained with the statistical and syntactic phrases. It is argued in section 4 that the use of cooccurrence improves the retrieval quality. The results seem to indicate that exactly the opposite is true. Upon analysis of the retrieval output it is discovered that the reason for this apparent turnabout is the dilution of nonrelevant concept vectors due to unmatched concepts. For many of the queries analyzed, there are one or more documents, highly correlated to that query, but nonrelevant, and which have a relatively large number of phrases which are not matched by the query. Because of the dilution effect which occurs when cooccurrence is not used, the correlations for these documents are lowered often below that of one of the relevant documents. The rank of the relevant document is thus raised by default although its correlation is often not altered. Consider for example the correlation of document 11 with query A4. With no phrases used, this nonrelevant document ranks sixth with a correlation of 0.24818. The document has 13 phrases which do not match the query. When retrieval is performed using these phrases without cooccurrence, the coefficient is reduced to 0.15599 and the rank lowered to ninth place. This allows one of the relevant documents to move ahead producing an apparent improvement in retrieval quality. When cooccurrence is used there are no phrase matches, the coefficient remains 0.24818, and the relevant document is not allowed to move up. Considering the entire set of 33 documents relevant to the test queries, the ranks of 16 are improved by the use of statistical phrases with no cooccurrence. However of these 16, only 7 actually move up in correlation coefficient. The remaining 9 lose in correlation but gain in rank due to the dilution and consequent lowering of nonrelevant documents. Ten of the 33 documents lose in both rank and coefficient, mostly

by dilution, while 7 remained fixed in rank. Of these 7, 5 are reduced in coefficient but by an amount insufficient to drop the rank. It is also discovered that most of the documents with a large number of phrases are not relevant to any test query. Thus the apparent improvement achieved when cooccurrence is not used is almost entirely due to the lowering in coefficient of certain nonrelevant documents. The fact that most of the documents with a large number of phrases are not relevant is not true in general; and the apparent improvement of experiment 2 over experiment 4 is clearly exceptional.

Attention is next focused on the fact that precision values obtained for statistical and syntactic phrase experiments with cooccurrence (experiments 4 and 5) fall below those achieved with no phrases at all. This can be understood by considering the experiment 4 results. Of the 33 relevant documents, this phrase process improves both the rank and correlation for 9; 5 are reduced in rank; while the remaining 19 are unchanged. Overall this seems to be an improvement, but this is not born out by the tabulated results in Fig. 7. The reason for this lies almost entirely with query B5. It has only one relevant document and the phrase process loweres its rank from second to fifth thus lowering its precision for all recall levels, from 0.5 to 0.2. This is a considerable decrease in precision, and since the values are averaged over only ten queries, the effect on the average is substantial. If precision values are taken for the nine other queries only, the values for the phrase processes exceed those for the no-phrase experiment for all recall levels. Thus with the exception of one rather unusual query, the results obtained by using the phrase process with cooccurrence (experiment 4 and, with similar reasoning, experiment 5) are slightly better than those obtained without phrases.

The tabulations in Fig. 7 indicate that results achieved by using the no-phrase-list method (experiment 6) are inferior to both the phrase list and no-phrase results. This is in part due to the method's inability to associate phrases with different constituent concepts. The inferior results can also be blamed on the very small number of cooccurrences. Of the more than 800 relations entered, only 28 cooccurrences between document and query are found. This very low number can be blamed, at least in part, on the queries. They are all quite short and contain very few phrases. The queries also tend to be quite general. Since retrieval is performed by concept matching and not by hierarchical expansion, general queries do not always produce the desired results. Of the 28 cooccurrences, only 5 occur between a query and one of its relevant documents. In the ten test queries, three have no cooccurrences at all, and their results are clearly not altered from the no-phrase case. Four queries have cooccurrences in nonrelevant documents only and these results are obviously lowered. The three remaining queries have cooccurrences in relevant documents; however, an improvement is realized in only one. Of the other two, one shows an improvement in correlation coefficient, but insufficient for a rank change, and the other has cooccurrences in nonrelevant documents which overshadowed any improvement. These results cast some doubt on the value of this method. However the present evidence is really inconclusive.

To summarize, it appears that contrary to the tabulated results, the phrase processes really do provide some improvement in retrieval performance. These improvements, however, are small and may not be worth the extra effort needed to achieve them. Before final judgment is passed on these tests, it must be determined whether the tests are fair. In the opening section of

this paper, the basic motivation for the use of phrases is stated to be the separation of highly correlated documents. A document collection thus has to provide such documents in order for a fair test to be conducted. A document-document correlation is performed on the ADI collection. The results are omitted here, but it is sufficient to say that with an average correlation of only 0.1, and a maximum of under 0.8, this collection is far too sparse to provide a conclusive test. The refinement added to the documents and queries by the use of syntax is not needed in such a sparse corpus and therefore substantial performance gains cannot be expected.

7. Conclusion

One general problem encountered in using phrases for retrieval concerns high frequency words. These words occur too frequently to be of any use in retrieval and they are therefore not included in the concept vectors. Certain of these words, however, can enter into significant phrases. The word "system" for example is one such high frequency word. By itself it carries little meaning, but in phrases such as "system programming" and "systems analysis", it is significant. But because "system" is a high frequency word, these phrases cannot be located. This problem will probably solve itself for larger and more general collections where the list of high frequency words is much more limited. While solving one problem, the larger collection creates another; that of the metalanguage; that is, words and phrases which convey no significant meaning as is seen in the underlined portion of the following sentence: In this paper the author discusses the methods of syntactic analysis. In the ADI collection metalanguage words are almost all included in the high frequency word list and thus eliminated.

This is impossible for larger collections and thus some means of automatically determining and eliminating metalanguage segments will have to be devised.

The results presented in this study seem rather discouraging. They do, however, point out a number of other areas that must be investigated in order to determine conclusively the value of phrases in retrieval. First, the experiments presented here must be retried using more dense collections. The amount of syntactic and semantic structure required in the concept vectors to achieve desired results seems to be related to the density of the collection. This relation should be investigated. A second area for further study is that of relevancy judgments. It is possible that significant improvement in retrieval quality cannot be achieved until some investigation is made into what makes a user deem a certain document relevant to a given query. And third, in order to implement syntactic or cooccurrence phrases into a usable system, an automated syntactic analyzer must be used. A number of such schemes exist including those by Hillman [1], Sager [2], and Baxendale [3]. None of these produce perfect results and further study must be made to improve this facility.

References

[1] Hillman, K., The Leader System, Proc. 1969 AFIPS Spring Joint Computer Conference, Thompson Book Co., Washington, May 1969.

[2] Sager, N., "Report on String Analysis", NSF Report, University of Pennsylvania, March 1966.

[3] Baxendale, P., "Documentation for an Economical Program for the Limited Parsing of English: Lexicon, Grammar, and Flowcharts" IBM Corporation Report RJ 386, August 1966.