VI. Suffix Dictionaries

E. M. Keen

1. Introduction

The use of suffix removal procedures as a simple method of vocabulary control is investigated with two types of suffix dictionaries. The need for vocabulary control and the desirability of synonym and partial synonym recognition are discussed in Section I. A suffix removal procedure has been incorporated into the SMART system from its inception, which has been known as the "null thesaurus", but is here described as the stem dictionary. A second type of dictionary recently tested is the "suffix 's' dictionary", since this offers the most basic language analysis method involving virtually no vocabulary control; as such, the suffix 's' method provides a convenient "base-line" from which dictionaries exerting greater control can be evaluated. A brief description of the two dictionaries will be given, together with retrieval performance comparisons and an analysis of the results.

2. Description of Suffix Dictionaries.

Both the suffix 's' and stem dictionaries are automatically generated, and the suffix removal procedure and collection look-up operations have been described elsewhere [1,2,3,4,5,6]. Briefly, the full suffix removal process (stem dictionary) is carried out in two stages: first, the construction of a dictionary of word stems, formed by applying a hand-made list of suffixes to a body of text; and second, by a look-up process which uses the dictionary of word stems plus certain spelling rules to reduce the documents texts to be stored in the system to word stems only. The suffix 's' dictionary is applied in the same manner, but in this case the only 'suffix' removed is the terminal 's', with the object of conflating singular and plural word forms.

Many of the considerations relating to the methods oı construction of the stem dictionary have been discussed by Salton and Lesk [7]. The comments made here relate to the extent to which the present dictionaries correctly conflate English word forms and so use the correct stems.

The conflation of singular and plural words is not perfectly achieved by terminal "s" removal, although over 70% success is obtained in the case of the Cran-1 aerodynamics terminology. The failures are due to well-known singular and plural forms such as "body" and "bodies", "axis" and "axes", "bureau" and "bureaux", "appendix" and "appendices", etc. Also, the terminal "s" does not always denote a plural form, and words like "bluntness" and "aerodynamics" have the "s" removed. This latter occurrence rarely affects retrieval, however, since a request and document both containing the word "bluntness" will match on the word without its terminal "s". It is possible to imagine a case of incorrect conflation, for example, the word "axe" could be incorrectly conflated with "axes", but such occurrences are extremely rare within the narrow subject fields under test.

The full suffix removal procedure incorporates spelling rules which correctly identify "bod" as the stem of both "body" and "bodies", and correctly conflate "hope", "hoped" and "hoping", as well as "hop", "hopped" and "hopping". There are some cases, however, where the correct stem is not recognized. For example, the words "computation", "computations" and "computational" are correctly conflated and given the same concept number as the look-up procedure, but a second group of similar words is given a second concept number including such words as "compute", "computed", "computers", "computer", and "computing".

A second example is the term "compressible", used in the aerodynamics literature, which is kept separately from "compressibility".

It appears that amendments to the automatic procedures used could solve at least some of these problems, and it is certain that for every such problem there are at least ten cases of correct conflation. Examination of the groups of words that are related by this conflating procedure suggests that the majority are helpful for document retrieval. A distinction between "computer" and "computing" is not believed to be useful, and preservation of the two forms is unlikely to be helpful to a requester. An exception to this situation may be furnished by the inclusion of a noun with the adjectival and verbal forms. Although the practice of using a "computer" is related to the "computer" itself, a request for documents describing one named computer may not perform well if documents describing computational procedures are highly matched with the request.

The performance results presented suggest that this type of unwelcome conflation is a contributing factor to the poor performance of the stem dictionary on the Cran-1 aerodynamics collection. The words "compressor" and "compressors", for example, are unhelpfully grouped with "compressible" and "compression", when notions such as "jet engine compressor", "compressible flow", and "compression buckling" are quite unrelated. Naturally any hand-produced dictionary, such as the thesaurus dictionaries described in section VII, can easily handle such conflation problems, but the claim for automatically generated dictionaries is that cases of failure are few enough to justify the large saving in effort of construction. This general claim seems to be potentially far better justified by the automatically generated thesaurus-type dictionaries produced by statistical association (see section VIII and appendix C), since hand construction of a stem dictionary would require little effort if an exhaustive concordance of the collection were available.

3. Retrieval Performance Results

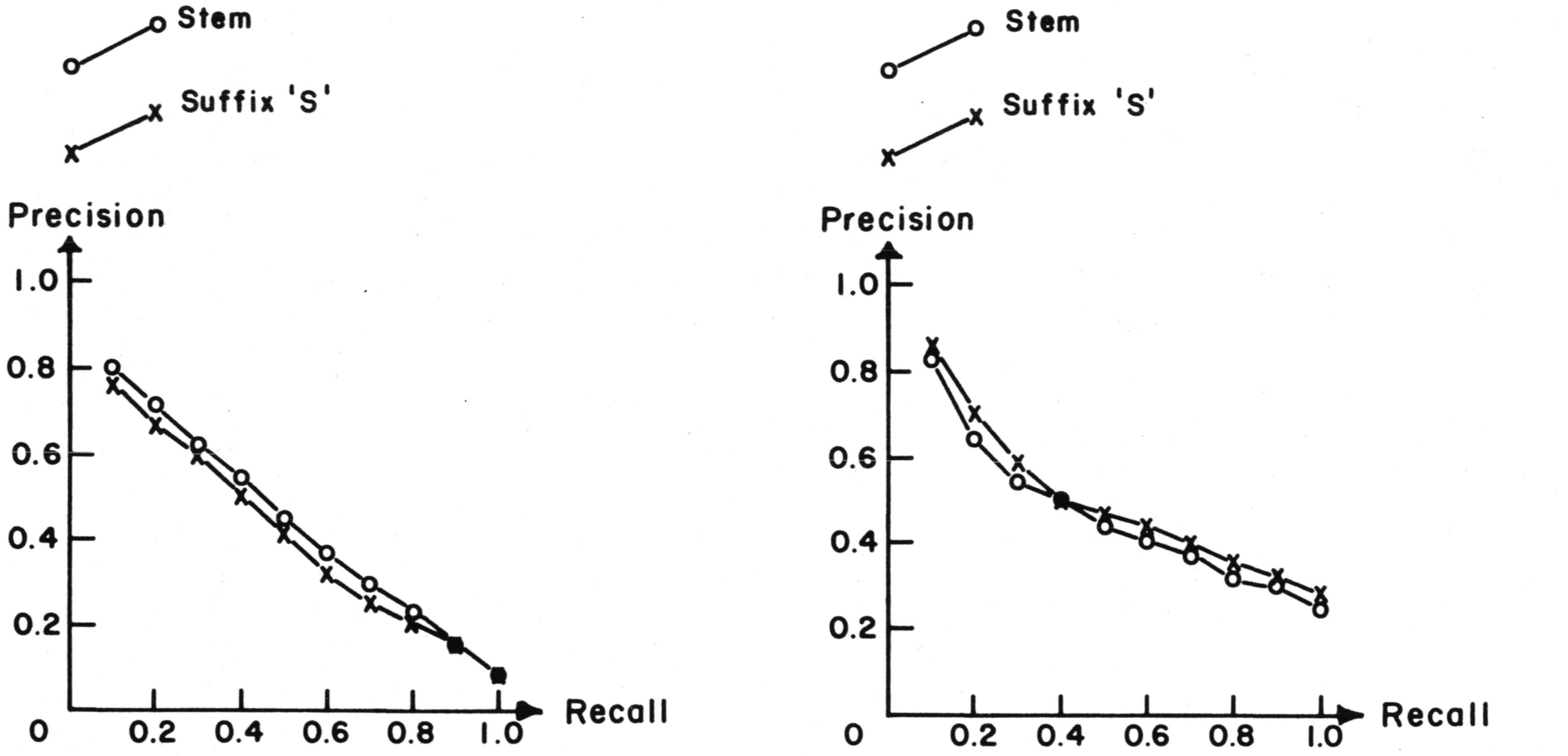
Comparisons of the suffix 's' and stem dictionaries are presented for the three document collections, using the normalized measures, precision versus recall graphs and data from individual requests. Figure 1 gives ten results using the normalized recall and precision measures. The ADI results include text, abstract and title results, and some results are displayed both for the ADI and IRE-3 collections with overlap correlation and logical vectors. All IRE-3 results and four of the six ADI results show the stem dictionary to have higher normalized values, although by quite small amounts. The single Cranfield result and the ADI text cosine and overlap logical runs show suffix 's' to be the superior dictionary.

Four results are given using precision versus recall graphs: IRE-3 Figure 2(a), Cran-1 Figure 2(b), ADI Abstracts Figure 3(a) and ADI Text Figure 3(b). These results confirm those in Figure 1, and the Cran-1 result is seen to favor suffix 's' over the whole range of the curve. To complete all the runs given in Figure 1 in terms of precision and recall, a table is given in Figure 4 that summarizes six more precision/recall plots not presented in detail, by recording the precision merit at three levels of recall. Some disagreement between these results and the normalized measures may be noted, and the reasons for this are discussed in section II. The cases of disagreement all consist of very small differences in merit between suffix 's' and stem, and all the more valuable comparisons which use the cosine correlation and numeric vectors display consistent results. The average performance measures show, therefore, that stem is superior to suffix 's' on the IRE-3 and ADI collections, and that suffix 's' is the better dictionary on the Cran-1 collection.

COLLECTION	INPUT AND MATCHING FUNCTION	EVALUATION MEASURE	STEM DICTIONARY	SUFFIX 'S' DICTIONARY
IRE-3 34 Requests	Abstract, Cosine Numeric	Normed. Recall Normed. Precision	.8954 .6746	.8817 .6484
	Abstract, Cosine Logical	Normed. Recall Normed. Precision	.8777 .6167	.8707 .6134
	Abstract, Overlap Logical	Normed. Recall Normed. Precision	.8725 .5829	.8408 .5611
CRAN-1 42 Requests	Abstract, Cosine Numeric	Normed. Recall Normed. Precision	.8644 .6704	.8717 .7018
ADI 35 Requests	Text, Cosine Numeric	Normed. Recall Normed. Precision	.7779 .5573	.7520 .5308
	Text, Cosine Logical	Normed. Recall Normed. Precision	.7695 .5248	.7768 .5462
	Text, Overlap Logical	Normed. Recall Normed. Precision	.7434 .4978	.7546 .5097
	Abstract, Cosine Numeric	Normed. Recall Normed. Precision	.7601 .5326	.7253 .4997
	Abstract, Cosine Logical	Normed. Recall Normed. Precision	.7546 .5221	.7296 .5044
	Title, Cosine Numeric	Normed. Recall Normed. Precision	.6722 .4537	.6435 .4209

Performance Results Comparing Stem and Suffix 's' Dictionaries for Ten Options on Three Collections, using Normalized Recall and Precision.

Figure 1.



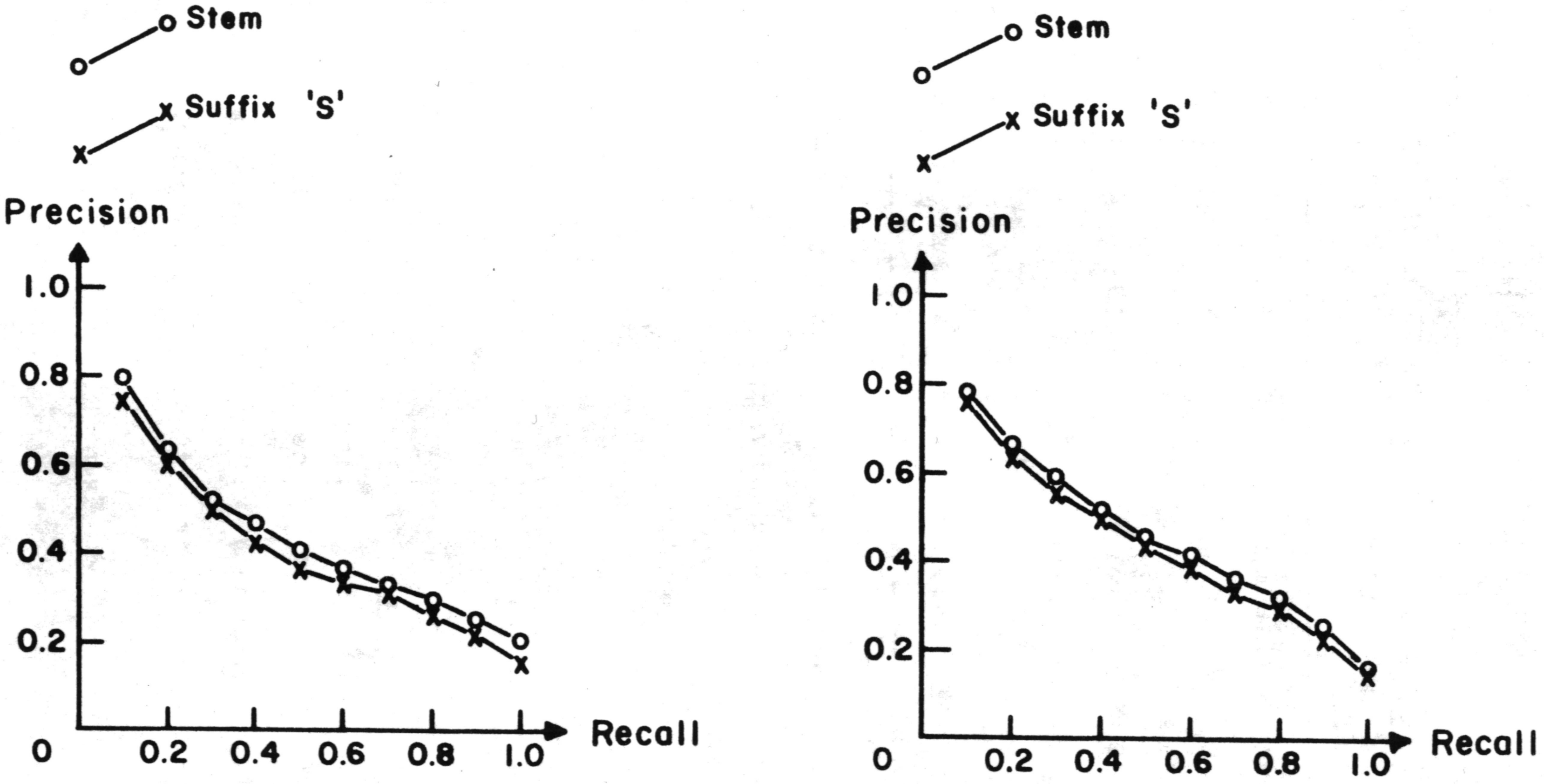
a) IRE-3, Abstracts, 34 Requests

b) Cran-1, Abstracts, 42 Requests

Performance Curves Comparing Stem and Suffix 's'

Dictionaries on Two Collections

Fig. 2.



a) ADI, Abstracts, 35 Requests

b) ADI, Text, 35 Requests

Performance Curves Comparing Stem and Suffix 's' Dictionaries

on the ADI Collection, using Abstracts and Full Text.

Fig. 3

COLLECTION	INPUT AND MATCHING FUNCTION	STEM(S) VERSUS SUFFIX 'S'(X)				
		PRECISION AT RECALL			NORMALIZED	
		.2	.5	.8	REC.	PRE.
IRE-3 34 Requests	Abstract, Cosine Logical	S	S	X	S	S
	Abstract, Overlap Logical	X	X	X	S	S
ADI 35 Requests	Text, Cosine Logical	X	X	S	X	X
	Text, Overlap Logical	X	S	S	X	X
	Abstract, Cosine Logical	S	S	S	S	S
	Title, Cosine Numeric	S	S	X	S	S

The merit of one dictionary over the other in these results is always by less than 0.05 precision, normalized recall and normalized precision.

Table Summarizing Six Precision Versus Recall
Plots not Presented, Comparing Stem and Suffix 's'
Dictionaries for Six Options on Two Collections.

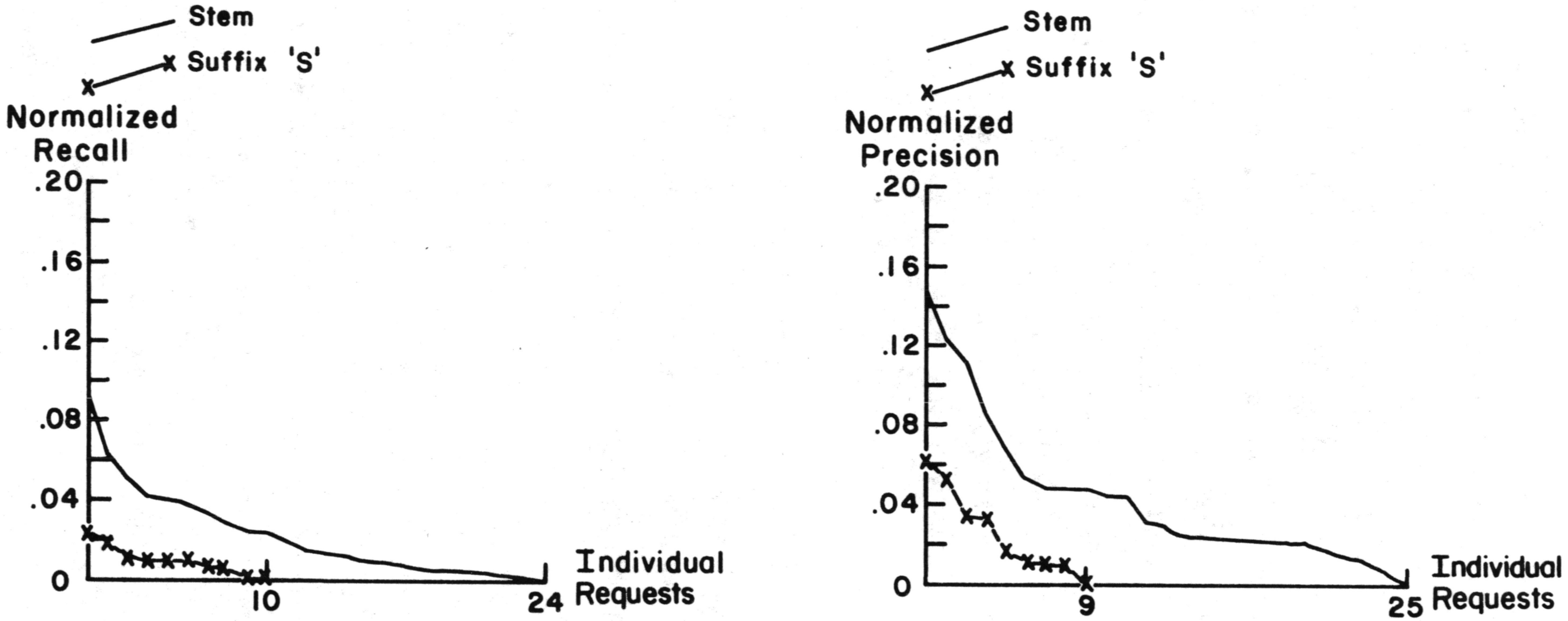
Figure 4.

These average results may be supplemented by the individual request data given in Figures 5, 6, 7 and 8. Using the normalized recall and precision measures as indicators of merit, it can be seen that 71% to 74% of the requests favor stem on IRE-3 (Figure 5), and 53% to 75% of the requests favor stem on ADI abstracts (Figure 7) and text (Figure 8). The Cran-1 result favoring suffix 's' is confirmed by figures relating to the individual request also, with 72% to 77% preferring suffix 's', ignoring those requests which have equal merit for both dictionaries. Each figure includes plots of both normalized recall and precision versus the individual requests. In the case of Cran-1 these plots show that suffix 's' is superior on the average because many of the requests favor suffix 's' by very small amounts. In the IRE-3 and ADI collections the stem dictionary displays some large changes in individual requests in its superiority over suffix 's'.

4. Performance Analyses

Two phenomena require explanation: firstly, the IRE and ADI runs involving logical vectors and overlap correlation which sometimes show suffix 's' superior to stem; and secondly, the superiority of suffix 's' on the Cran-1 collection.

The first phenomenon is less important than the second, because logical and overlap runs are inferior to cosine numeric runs in any case. Cases where suffix 's' is better than stem must be caused by circumstances of the type considered in part 2, where full suffix removal conflates some words that match with non-relevant documents and thus adversely affect performance. It was noted in section III that the use of numeric vectors (weighted) gives a clear advantage over logical vectors when a dictionary is in use that includes a reasonably large amount of mapping (i.e., it



24 Requests better on Stem (70.6%)

10 Requests better on Suffix 's' (29.4%)

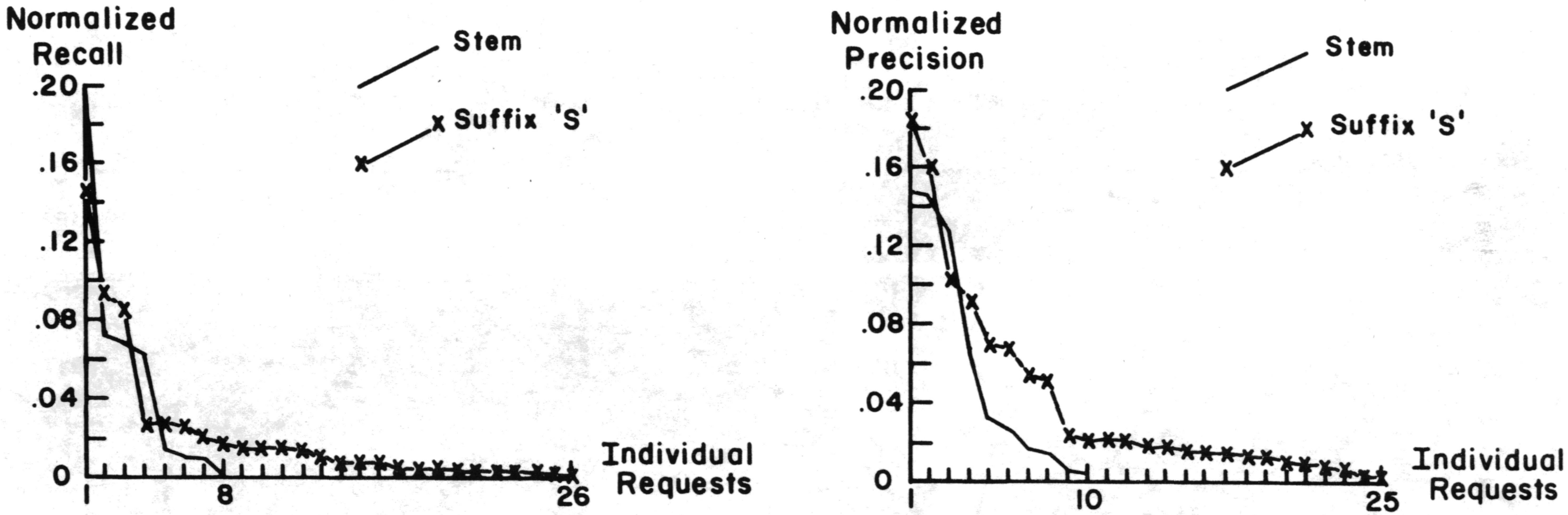
25 Requests better on Stem (73.5%)

9 Requests better on Suffix 's' (26.5%)

IRE-3, 34 Requests

Graphs of the Magnitudes of the Differences in Individual Requests Comparing Stem and Suffix 's' Dictionaries, using Normalized Recall and Precision

Fig. 5.



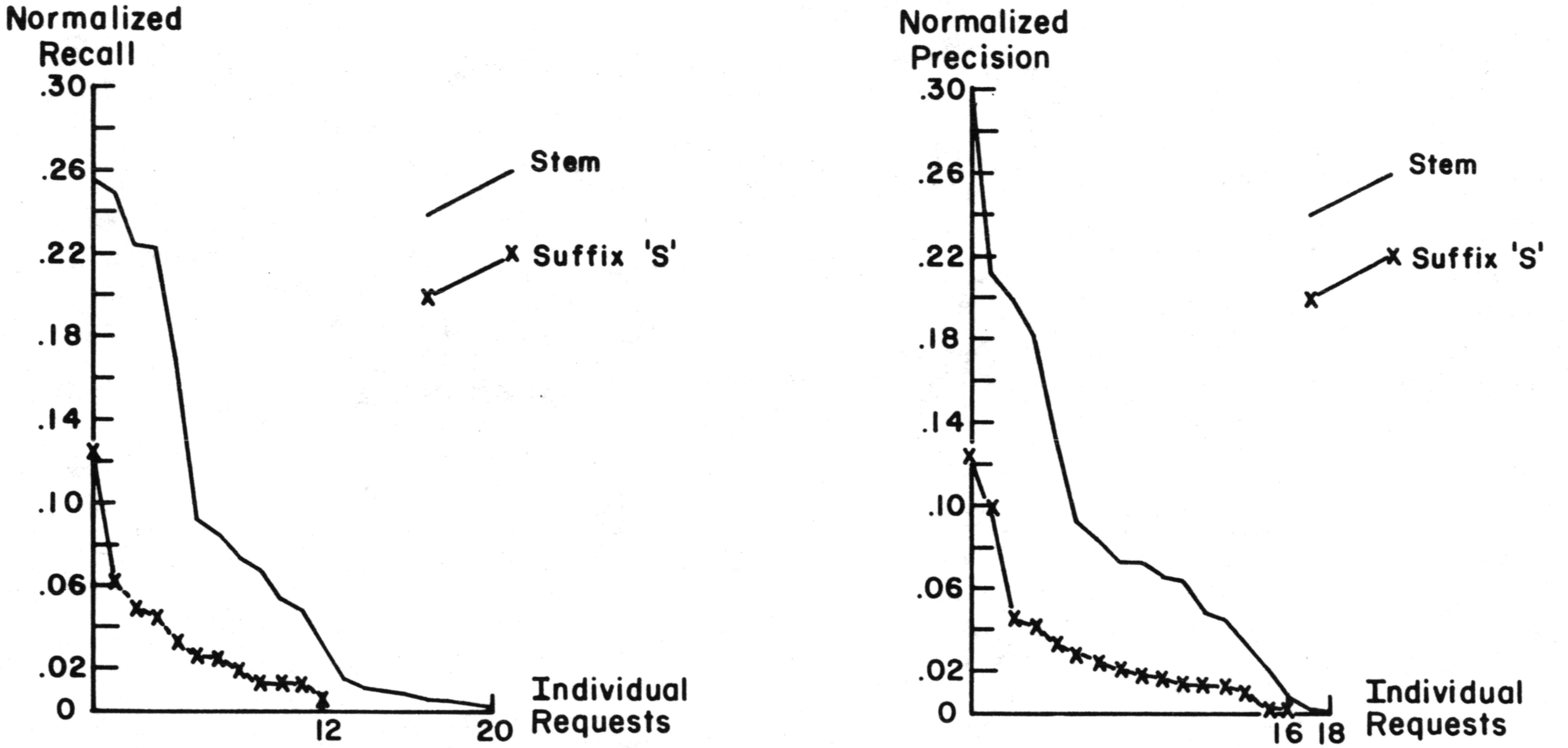
26 Requests better on Suffix 's' (76.5%)
8 Requests better on Stem (23.5%)
8 Requests equal on both

25 Requests better on Suffix 's' (71.4%)
10 Requests better on Stem (28.6%)
7 Requests equal on both

CRAN-1, 42 Requests

Graphs of the Magnitudes of the Differences in Individual Requests, Comparing Stem and Suffix 's' Dictionaries, using Normalized Recall and Precision

Fig. 6.

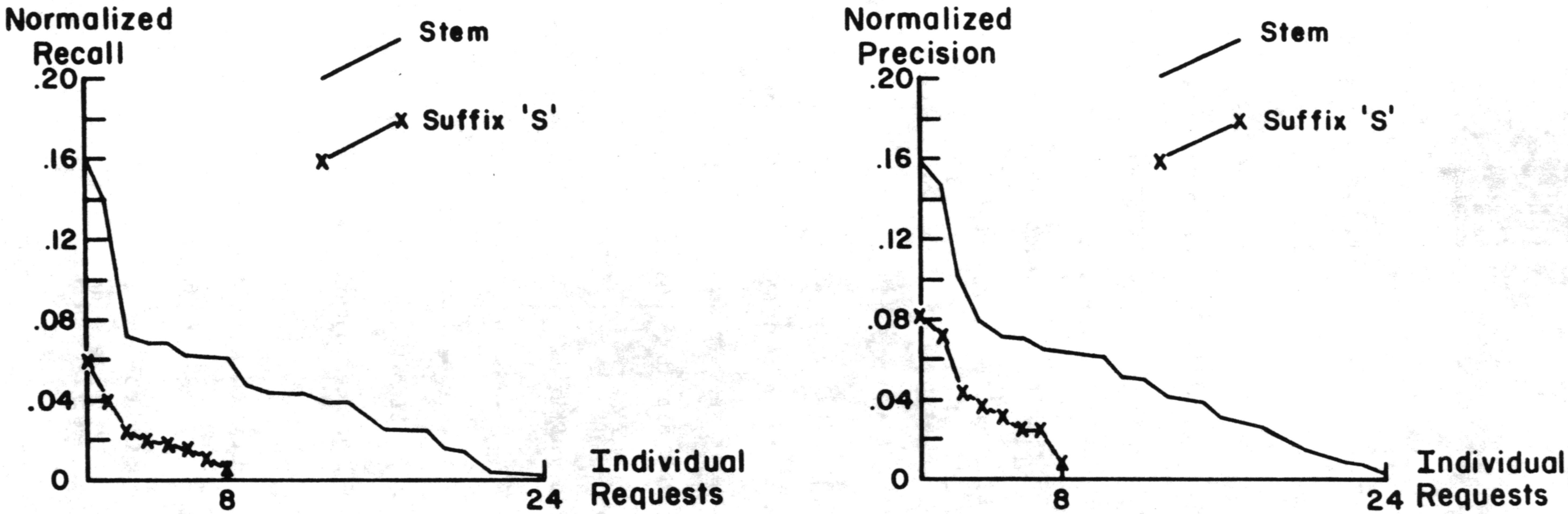


20 Requests better on Stem (62.5%)
12 Requests better on Suffix 's' (37.5%)
3 Requests equal on both

18 Requests better on Stem (52.9%)
16 Requests better on Suffix 's' (47.1%)
1 Request equal on both

ADI, Abstracts, 35 Requests

Graphs of the Magnitudes of the Differences in Individual Requests, Comparing Stem and Suffix 's' Dictionaries, using Normalized Recall and Precision.
Fig. 7.



24 Requests better on Stem (75%)
8 Requests better on Suffix 's' (25%)
3 Requests equal on both

24 Requests better on Stem (75%)
8 Requests better on Suffix 's' (25%)
3 Requests equal on both

ADI, Text, 35 Requests

Graphs of the Magnitude of the Differences in Individual Requests, Comparing Stem and Suffix 's' Dictionaries, using Normalized Recall and Precision

Fig. 8

conflates many words), and a similar, but unexplained, relationship is noted when the use of cosine is compared to overlap. From a strictly experimental viewpoint dictionaries such as suffix 's' and stem should be compared without the addition of weighting procedures and cosine, in order that the dictionary mapping characteristics may be tested alone. In this case, the overlap logical results show that stem and suffix 's' dictionaries perform very similarly, and therefore within the context of the requests and relevance decisions in use, no advantage should be gained from full suffix recognition as performed automatically. This finding is in accordance with the general conclusions of the second Aslib-Cranfield Project [8], although in those results the nearest equivalent to the stem dictionary does perform a little better than suffix 's'.

However, a more practical conclusion in the case of SMART is that stem is the superior dictionary on the IRE-3 and ADI collections, since the cosine correlation and numeric vectors have clearly been proved to be advantageous, and would be advocated for use in any operational version of SMART.

The superiority of suffix 's' on Cran-1 is one of several instances where the Cran-1 result differs from the other collections. In the case of Cran-1 the difference in word mapping between suffix 's' and stem is less marked than in the other collections, since Figure 9 shows that the Cran-1 stem dictionary includes 83% of the concept classes contained in suffix 's', whereas the IRE-3 and ADI stem dictionaries are based on more mapping characteristics, including only 76% and 74% of suffix 's', respectively. As expected, this affects the match with requests and documents, since Figure 10 shows that at a cosine correlation cut-off of 0.35, the stem dictionary in Cran-1 does not retrieve so many additional documents over suffix 's' than is true for the other collections.

COLLECTION AND SIZE	CONCEPT CLASSES IN DICTIONARIES		
	SUFFIX 'S'	STEM	STEM AS % OF SUFFIX 'S'
IRE-3, 780	5,150	3,933	76.4%
ADI Text, 82	7,615	5,606	73.6%
Cran-1, 200	3,181	2,628	82.6%

Comparisons of Stem and Suffix 's' Dictionary Sizes on the Three Collections.

Figure 9.

COLLECTION	NUMBER OF DOCUMENTS ABOVE CORRELATION 0.35		INCREASE IN NUMBER OF DOCUMENTS ABOVE CORR. 0.35 STEM AS PERCENTAGE OF SUFFIX 'S'
	SUFFIX 'S'	STEM	
IRE-3 (34 Requests)	157	222	41.4%
ADI(Abstracts) (35 Requests)	40	55	37.5%
ADI (Text) (35 Requests)	42	61	45.2%
CRAN-1 (43 Requests)	42	54	25.6%

Comparisons of Numbers of Documents with a Cosine Numeric Correlation above 0.35 on Stem and Suffix 's' Dictionaries, using Four Results from Three Collections.

Figure 10.

This data shows that full suffixing (stem dictionary) does not affect so many aerodynamics words as it does in computer science and documentation, thus giving the Cran-1 collection less of an opportunity for a change in retrieval performance. Further explanations can be given by observing individual request performance for the seven requests that have performance changes greater than 0.05 normalized recall (see Figure 6); four of these requests are better on stem, and three are better on suffix 's'. An analysis of the three requests that favor suffix 's' reveals certain test problems, connected mainly with hyphenation and keypunch errors, and in fact the request/relevant document match is in all cases very weak. The many requests that favor suffix 's' by a trivial amount (Figure 6) are typified by request Q269, details of which appear in Figure 11. As Figure 11 shows, the stem "compress" incorrectly matches the request word "compressor" with a frequently used word "compressible" in two non-relevant documents, so that the stem dictionary has an inferior performance because relevant document 1590 receives a rank position below the two non-relevant documents. In case the matching words of non-relevant document 1984 appear to put into question the relevance decision (the abstract includes the topic of "choked flow in an impeller inlet"), the title makes it clear that it is a "centrifugal impeller", and the matching word "axial" is a spurious match from the phrase "axial symmetry".

The ranks of the seven documents relevant to request Q190 are given in Figure 12, because the changes in rank position observed are typical of what happens to the averages. The highest ranked relevant document remains unchanged at position 1; thus the high precision end of the curve for this request will remain unchanged. The next three relevant items are ranked better by suffix 's', but the final three relevant are ranked better by stem. This result is seen to give a greater superiority to suffix 's' in

Request Q269 (Cran-1 Collection)

Has a criterion been established for determining the axial compressor choking line?

Suffix 's' Dictionary

Rank	Document	Cosine Numeric Correlation	Matching Words and Weights
2	1591 Relevant	.2817	Compressor(3) Line(2)
5	1590 Relevant	.1375	Axial(1) Compressor(2)
7	1967 Non-Relevant	.0922	Axial(3)
20	1984 Non-Relevant	.0301	Axial(1)

Stem Dictionary

Rank	Document	Cosine Numeric Correlation	Matching Words and Weights
1	1591 Relevant	.4234	Compressor(4) Choke(2) Line(2)
5	1967 Non-Relevant	.1744	Axial(3) Compressible(2) Determined(1)
6	1984 Non-Relevant	.1532	Axial(1) Compressible(2) Choked(1) Determined(1)
7	1590 Relevant	.1365	Axial(1) Compressor(2)

Individual Aerodynamics Request Q269 Showing Superior Performance Obtained with Suffix 's' Dictionary for Relevant Document 1590.

Figure 11.

Request Q190 (Cran-1 Collection)

Ranks of the 7 Relevant Documents

Suffix 's'		Stem	
Rank	Relevant Document	Rank	Relevant Document
1	987	1	987
3	988	7	988
6	989	17	989
20	984	21	984
30	985	44	985
60	990	53	990
81	986	61	986

Nor. Rec. = 0.8719	Nor. Rec. = 0.8697
Nor. Pre. = 0.6745	Nor. Pre. = 0.6072

Individual Aerodynamics Request Q190 Comparing Suffix 's' and Stem Dictionary Performance.

Figure 12.

normalized precision than normalized recall, as the averages show (Figure 1). But the precision/recall curve is little affected by dictionary change when averaged over all requests, as Figure 2(b) shows.

A definite conclusion must await an investigation into the effect of changes in subject language and the effects of differing methods of request and relevance decision preparation, since both factors are involved in a comparison of Cran-1 with the other two collections. Meanwhile, the evidence presented does point to a difference in language characteristics, and tests on the larger Cran-2 collection of 1400 documents will shed more light on this.

5. Conclusions

The comparison of the two suffixing dictionaries shows stem to be superior on the IRE-3 and ADI collections, and suffix 's' to be superior on the Cran-1 collection. All differences between dictionaries are small, and the use of overlap correlation and logical vectors on the IRE-3 and ADI collections lessen the superiority of stem; however, the cosine numeric result is to be preferred to these procedures. The aerodynamics terminology appears to offer less opportunity for word conflation than the computer science and documentation terminologies; this remains the primary explanation so far discovered for the Cran-1 result.

Every indication shows that the suffixing dictionaries provide a convenient and valid base-line from which further dictionaries of the thesaurus type can be evaluated. However, the use of some type of suffixing dictionary does provide a good retrieval tool in its own right. Such dictionaries should be considered both as tools that can be constructed with a minimum of effort in the absence of a thesaurus dictionary, and also as probable candidates for inclusion in systems in which a series of several dictionaries are provided from which a pre-search choice can be made. If the latter reason for inclusion of a stem dictionary is valid, then it would seem that a hand edited version, which would require little human effort, would probably overcome many of the detailed deficiencies that have been described.

References

[1] C. Harris, Dictionary and Hierarchy Formation, Information Storage and Retrieval, Report ISR-7, to the National Science Foundation, Section III, Harvard Computation Laboratory, June 1964.

[2] M. E. Lesk and T. Evslin, Housekeeping Routines, Information Storage and Retrieval, Report ISR-7, to the National Science Foundation, Section XI, Harvard Computation Laboratory, June 1964.

[3] C. Harris, Dictionary Construction and Updating, Information Storage and Retrieval, Report ISR-8, to the National Science Foundation, Section VII, Harvard Computation Laboratory, December 1964.

[4] M. Cane, The Dictionary Lookup System, Information Storage and Retrieval, Report ISR-9, to the National Science Foundation, Section V, Harvard Computation Laboratory, August 1965.

[5] M. Cane, The Dictionary Setup Procedures, Information Storage and Retrieval, Report ISR-9, to the National Science Foundation, Section VI, Harvard Computation Laboratory, August 1965.

[6] M. E. Lesk, Operating Instructions for the SMART Text Processing and Document Retrieval System, Information Storage and Retrieval, Report ISR-11, to the National Science Foundation, Section II, Department of Computer Science, Cornell University, June 1966.

[7] G. Salton and M. E. Lesk, Information Analysis and Dictionary Constuction, Information Storage and Retrieval, Report ISR-11, to the National Science Foundation, Section IV, Department of Computer Science, Cornell University, June 1966.

[8] C. Cleverdon and M. Keen, Factors Determining the Performance of Indexing Systems, Volume 2, Test Results, Aslib Cranfield Research Project, Cranfield, 1966.