VII. Thesaurus, Phrase and Hierarchy Dictionaries

E. M. Keen

1. Introduction

The suffix removal procedures described in Section VI provide synonym control only when identical word stems are involved; any comprehensive synonym and partial synonym recognition requires a procedure that groups words according to synonymy irrespective of word spelling. For this reason, the use of dictionaries of the thesaurus type is being investigated, as well as the use of phrases rather than single words, and also the use of word relations as specified by hierarchical arrangements. The construction characteristics of several dictionaries are discussed in the present section, before retrieval runs are presented, using retrieval results for three document collections.

2. Description of Thesaurus Dictionaries

Seven thesaurus dictionaries are currently available, and each is referred to as follows:

1. IRE-3 Thesaurus-2. Known also as the "Harris 2" thesaurus, this handmade dictionary was originally constructed for use specifically with the IRE-1 collection.
2. IRE-3 Thesaurus-3. Known also as the "Harris 3" thesaurus, this handmade dictionary was constructed for use with any collection of computer science documents, and was first tested on the IRE-2 collection.
3. CRAN-1 Thesaurus-1. Known also as the "Old Quasi-Synonym" dictionary, this is a modified manually-constructed version of

the quasi-synonym list used in the Aslib Cranfield Project [1].

4. CRAN-1 Thesaurus-2. Known also as the "New Quasi-Synonym" dictionary. This dictionary was constructed by rearranging the word groups and incorporating additional words into the old quasi-synonym dictionary, using five specified rules for dictionary construction [2].

5. CRAN-1 Thesaurus-3. Known also as the "Revised New Quasi-Synonym" dictionary, this revision was made primarily to permit processing of the larger CRAN-2 collection, and involved also some small changes in grouping of the words.

6. ADI Thesaurus-1. Known also as a "regular thesaurus", this handmade dictionary was constructed for use with the full text ADI collection.

7. ADI Thesaurus-SA1. Known also as the "Hastie" dictionary, this represents an attempt to use the semi-automatic procedures suggested in [2].

Some discussion of the construction expertise that has been gained by experience is contained in a number of previous reports. [2,3,4,5,6,7,8] Synonyms and other less closely related words are grouped subjectively in the case of manually constructed dictionaries, and the effectiveness of a particular dictionary can be determined by comparing the resulting retrieval performance for a set of search requests with the performance obtained with a stem dictionary. The main objective data that can be derived from a thesaurus construction algorithm is the amount of word grouping, measured by the average number of distinct natural language text words that are grouped into a thesaurus concept, and also the amount of overlap or ambiguity, measured by the number of words that appear in more than one concept group. This data is given for the seven dictionaries in Fig. 1.

Ignoring the semi-automatic "Hastie" ADI Thesaurus-SA1 and the Cran-1 Thesaurus-1 (made without use of the construction rules), the dictionaries average 594 concepts each, with 10.1 text words grouped into each concept.

Some sample excerpts from three dictionaries illustrating the grouping of similar terms in the context of three collections used are given in Fig. 2. It may be noted that a topic such as "Algebra" or "Calculate" is grouped only with almost synonymous terms (if any exist) when these topics are central to the collection in use, but a broader grouping is used when these topics are more peripheral to the subject field of the collection. Hyphenated word pairs are normally treated as a single word and usually put with the group most closely associated; for example "computing-machine" is put in the group which includes "computer" rather than the group including "machine". The need to group single words creates problems of ambiguity that are only partially solved by putting such words into more than one group. The word "factor", for example, may need to be grouped with "coefficient" as well as with "parameter" and "variable", but an incoming request containing"factor" then maps into several thesaurus groups, and only a decrease in weight resulting from the multiple mapping is then available to attempt to minimize the effect of the unwanted association. Some suggestions for further studies on dictionary construction are given in part 8.

3. Description of Phrase Dictionaries

Since the thesaurus dictionaries contain single words only, some kind of phrase processing is a reasonable alternative for dictionary construction.

Collection and Dictionary	Distinct Text Words	Concepts in Thesaurus	Average Words Per Concepts	Word Stems Appearing in More Than One Thesaurus Concept	
				Total	Percent
IRE-3 Thesaurus-2	*5,477	511	10.72	451	8.2%
IRE-3 Thesaurus-3	*5,477	686	7.98	159	2.9%
CRAN-1 Thesaurus-1	3,291	377	8.73	155	4.7%
CRAN-1 Thesaurus-2	3,291	495	6.65	389	11.8%
† CRAN-2 Thesaurus-3	*7,449	736	10.1	78	1.1%
ADI Thesaurus-1	8,099	541	14.97	54	0.7%
ADI Thesaurus-SA1	8,099	289	28.02	416	5.1%

* Estimated Values

† Data for Cran-1 Use of this Thesaurus are not available.

Grouping Characteristics of Seven Thesaurus Dictionaries

Fig. 1

IRE-3 Thesaurus-3	CRAN-1 Thesaurus-3	ADI Thesaurus-1
(Computer Science)	(Aerodynamics and Aeronautical Engineering)	(Documentation)
605 Calculate Compute 13 Evaluate Interpolate Plot Recompute 148 Add Sum 7 Algebra 376 Arithmetic 116 Mathematic 601 Computer Data-processor Electronic-computer	20 Algebra Arithmetic Calculate Compute Derivation Mathemat Newly-computed Numerical 601 Extrapolate Interpolate Quadrature 304 Analog Analogue Computer Computing Machine Digital-computer Digital IBM-704 IBM Univac	350 Calculate Compute Interpolate Sum 428 Mathemat 2 Computer-based Computer

Sample Excerpts from Three Thesaurus Dictionaries

Fig. 2

Descriptions of the methods used by SMART have previously appeared in [2,3,5,6,8,9,10,11]. No studies have yet been made of full-scale phrase recognition, and the "statistical phrase" technique used is intended only to remove cases of single word ambiguity. For example, a hypothetical medical request on "swine fever in New Guinea" will be quite strongly matched, using a thesaurus, with a document dealing with "diseases of the guinea pig". The use of a phrase dictionary containing "New Guinea" would give strong weight to the occurrence of both "New" and "Guinea" in a sentence, and thus the spurious match with "Guinea" in the sense of "guinea pig" would receive less weight by comparison.

The phrase dictionaries tested are handmade, and are based on the thesaurus groups. Phrase recognition takes place if the two or more component words (thesaurus concept numbers) appear in the same sentence; no specific word order position or syntactical relation is demanded. Phrases are used in retrieval as an addition to the thesaurus dictionary; thus, when a phrase occurs, a new concept identifier is added to the thesaurus concepts already assigned to the request or document, or the weight of an existing concept identifier is increased.

These procedures may be clarified by the excerpt from a thesaurus and phrase dictionary given in Fig. 3. The phrase made up from the thesaurus groups containing "axial" and "symmetry" is of value because the word "axial" is more commonly to be found in conjunction with "compressor"; thus, without phrase processing, any document dealing with "axial compressors" that also contains a concept identifier such as "regular" or "uniform" could be matched with a request for "axial symmetry". The addition of phrase processing in this example does not prevent such a

376	Emiss	185	Axes
	Emit		Axial-force
	Radiate		Axial
			Coaxial
388	Effect		X-axis
	Phenomen		etc.
		265	Regular
423	Ultra-violet		Symmetric
	Ultraviolet		Symmetry
	X-ray		Uniform
474	Solar	264	Axially-symmetric
			Axiallysymmetric
533	423 376		Axi-symmetric
	474 376		Axisymmetric
	474 388		185 265

Excerpts from the Cran-2 Thesaurus-3 Dictionary with Phrases Showing the Grouping to Recognize the Phrases "Axial Symmetry", "Ultraviolet Radiation", "Solar Emission" and "Solar Effect"

Fig. 3

spurious match, but it gives considerably greater weight to a correct phrase match. Fig. 3 also shows that some new synonymous concepts are produced by phrases, since the related notions of "ultraviolet radiation" and "solar emission" are not properly related in the thesaurus dictionary alone.

It is a simple matter to invent examples where this kind of phrase processing can lead to spurious matches, both because thesaurus concept groups are used as phrase components, and because within-sentence occurrence is the only criterion for recognizing a phrase. However, the document collections in use deal with quite restricted subject areas, and an examination shows that around 90% of the phrases recognized are either completely correct or at least legitimate for retrieval purposes. An example of a legitimate, but not strictly correct, phrase is the recognition of "boundary conditions" in a sentence containing the phrases "boundary layer" and "surface conditions".

A more reasonable criticism of the phrase procedures is the fact that too few phrases are listed in the dictionaries, as the data in Fig. 4 shows. However, if more complete phrase recognition procedures were used, the size of the phrase dictionaries would vastly exceed the size of the present thesaurus dictionaries, and the co-occurrence recognition procedures to be used would probably have to become more sophisticated than is presently the case.

4. Description of Hierarchy Dictionaries

The use of hierarchies provide formal relationships used in processing

Collection and Dictionary	Number of Phrases
IRE-3 Thesaurus-2	93
IRE-3 Thesaurus-3	374
CRAN-2 Thesaurus-3	309
ADI Thesaurus-1	247

Data on Phrase Dictionaries

Fig. 4

search requests is quite commonplace in document retrieval. In addition, the words grouped in the thesaurus dictionary may display hierarchical relationships; for example, concept 22 of the Cran-2 Thesaurus-3 groups both "algebra" and "arithmetic" with the generic notion of "mathematics" (Fig. 2). Hierarchy dictionaries tested have been constructed by structuring the thesaurus concepts themselves, rather than by going back to the separate words or word stems. Hierarchies have been manually constructed only for the IRE Computer Science collection, and descriptions of the methods used in their construction have appeared in [2,3,5,8,13,14]. Discussion and evaluation of procedures for automatically producing hierarchies by co-occurrence statistics is also not considered here (see [2,15]).

5. Retrieval Performance Results

A) Thesaurus Dictionaries

Performance comparisons are normally made between the stem and thesaurus dictionaries, and a series of comparisons using normalized recall and precision are given in Figs. 5, 6, and 7. The results in Fig. 5 are all based on the cosine numeric matching function, and it may be seen that even with different document input lengths,the thesaurus dictionaries are nearly always superior to stem. Reasonable explanations can be found for two main exceptions, since the Cran-1 Thesaurus-1 was made without the use of any of the construction rules; furthermore, it was based on the indexing only, omitting many words which appeared in the abstracts. The second exception is the ADI "Hastie" Thesaurus-SA1 which was made by semi-automatic procedures and was known to contain unsatisfactory groupings. Figs. 6 and 7 give, respectively, some results based on the Cosine Logical and Overlap Logical

No.	Collection	Input and Type of Thesaurus	Evaluation Measure	Stem Dictionary	Thesaurus Dictionary
1	IRE-3 34 Requests	Abstract (Thesaurus-2)	Normed. Recall Normed. Precision	.8954 .6746	.9191 .7072
2		Abstract (Thesaurus-3)	Normed. Recall Normed. Precision	.8954 .6746	.9268 .7382
3		Title (Thesaurus-2)	Normed. Recall Normed. Precision	.8145 .5547	.8436 .5945
4		Title (Thesaurus-3)	Normed. Recall Normed. Precision	.8145 .5547	.8430 .6068
8	Cran-1 42 Requests	Abstract (Thesaurus-1)	Normed. Recall Normed. Precision	.8644 .6704	.8602 .6319
9		Abstract (Thesaurus-2)	Normed. Recall Normed. Precision	.8644 .6704	.8864 .6864
10		Abstract (Thesaurus-3)	Normed. Recall Normed. Precision	.8644 .6704	.8837 .6952
11		Title (Thesaurus-3)	Normed. Recall Normed. Precision	.8112 .6185	.8374 .6420
12		Indexing (Thesaurus-1)	Normed. Recall Normed. Precision	.8897 .6831	.8629 .6335
13		Indexing (Thesaurus-2)	Normed. Recall Normed. Precision	.8897 .6831	.8992 .7094
16	ADI 35 Requests	Text (Thesaurus-1)	Normed. Recall Normed. Precision	.7779 .5573	.8206 .6273
17		Text (Thesaurus-SA1)	Normed. Recall Normed. Precision	.7779 .5573	.7774 .5441
18		Abstract (Thesaurus-1)	Normed. Recall Normed. Precision	.7601 .5326	.8016 .6069
19		Abstract (Thesaurus-SA1)	Normed. Recall Normed. Precision	.7601 .5326	.7548 .5190
20		Title (Thesaurus-1)	Normed. Recall Normed. Precision	.6722 .4537	.7324 .5462
21		Title (Thesaurus-SA1)	Normed. Recall Normed. Precision	.6722 .4537	.6877 .4649

Performance Results Comparing Stem and Thesaurus Dictionaries for Sixteen Results using Cosine Numeric on three Collections, and Normalized Recall and Precision

Fig. 5

No.	Collection	Input and Type of Thesaurus	Evaluation Measure	Stem Dictionary (cosine log.)	Thesaurus Dictionary (cosine log.)
5	IRE-3 34 Requests	Abstract (Thesaurus-3)	Normed. Recall Normed. Precision	.8777 .6167	.9067 .6574
14	Cran-1 42 Requests	Abstract (Thesaurus-3)	Normed. Recall Normed. Precision	.8397 .6377	.8729 .6936
22	ADI 35 Requests	Text (Thesaurus-3)	Normed. Recall Normed. Precision	.7695 .5248	.7819 .5092
23		Text (Thesaurus-SA1)	Normed. Recall Normed. Precision	.7695 .5248	.6884 .4332
24		Abstract (Thesaurus-1)	Normed. Recall Normed. Precision	.7546 .5221	.8043 .5823

Performance Results Comparing Stem and Thesaurus Dictionaries for Five Results using Cosine Logical on Three Collections, and Normalized Recall and Precision

Fig. 6

No.	Collection	Input and Type of Thesaurus	Evaluation Measure	Stem Dictionary (Overlap Logical)	Thesaurus Dictionary (Overlap Logical)
6	IRE-3 34 Requests	Abstract (Thesaurus-2)	Normed. Recall Normed. Precision	.8725 .5829	.8840 .5775
7		Abstract (Thesaurus-3)	Normed. Recall Normed. Precision	.8725 .5829	.8974 .6041
15	Cran-1 42 Requests	Abstract (Thesaurus-3)	Normed. Recall Normed. Precision	.8237 .5830	.8535 .6251
25	ADI 35 Requests	Text (Thesaurus-1)	Normed. Recall Normed. Precision	.7434 .4978	.7386 .4350
26		Text (Thesaurus-SA1)	Normed. Recall Normed. Precision	.7434 .4978	.6589 .3602
27		Abstract (Thesaurus-1)	Normed. Recall Normed. Precision	.7423 .4904	.7830 .5257

Performance Results Comparing Stem and Thesaurus Dictionaries for Six Results Using Overlap Logical on Three Collections, and Normalized Recall and Precision

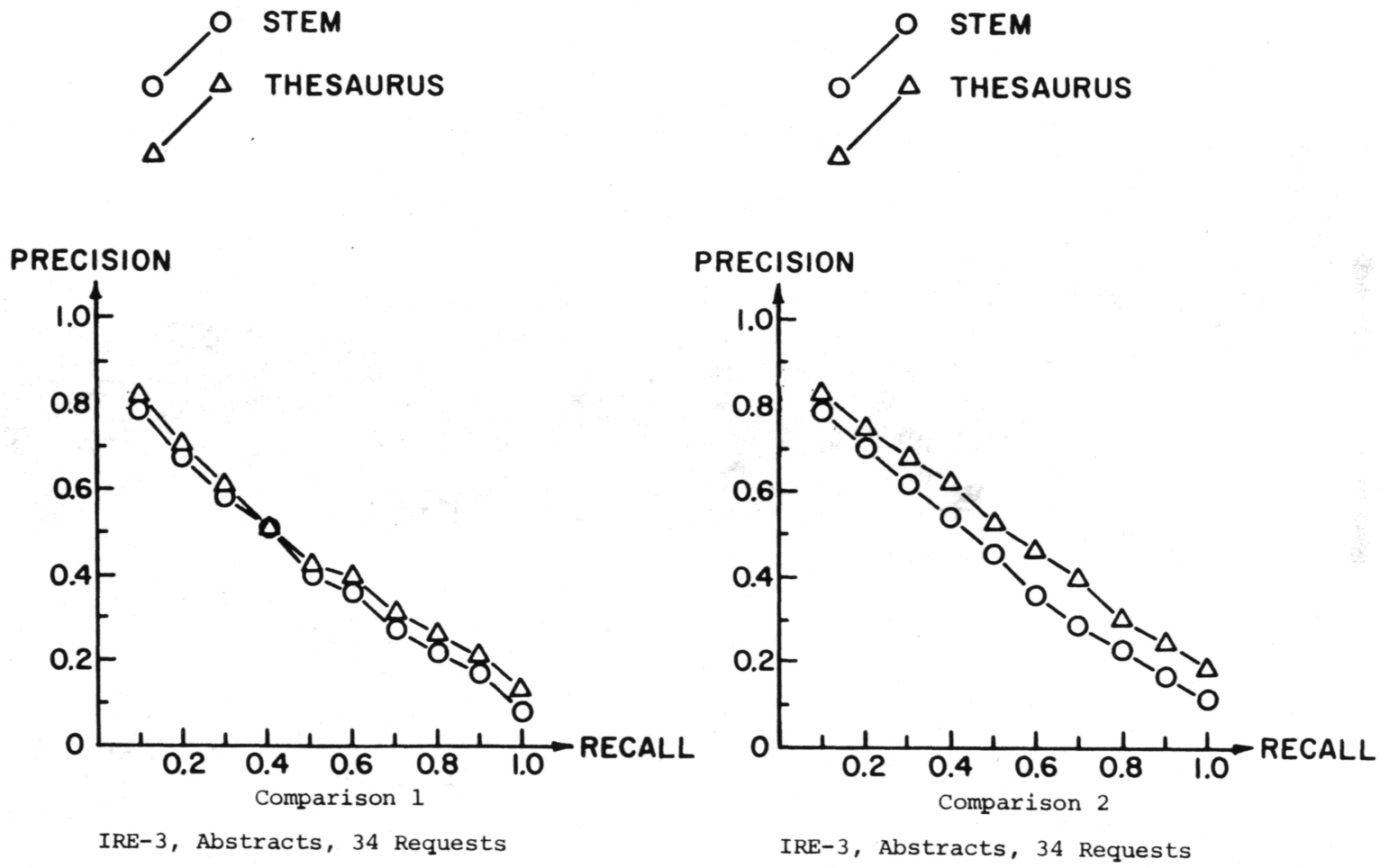
Fig. 7

matching functions. These again display a superiority for thesaurus in the expected cases, except for the "ADI Text Thesaurus-1 Overlap Logical" result.

Precision versus recall graphs are used to repeat the most important comparisons, with IRE-3 in Fig. 8, Cran-1 Abstracts in Figs. 9 and 10, Cran-1 Indexing in Fig. 11, and ADI Abstracts and Text in Fig. 12. Thesaurus works better than stem in all cases except in that of the first Cranfield version. Fig. 10 compares thesaurus with suffix 's', since on Cran-1, stem is not as good as suffix 's' (both at the high precision end and between 0.65 and 0.85 recall suffix 's' is superior). The figures do show that the thesaurus dictionaries are superior by a much greater amount for both IRE-3 and ADI than for Cran-1.

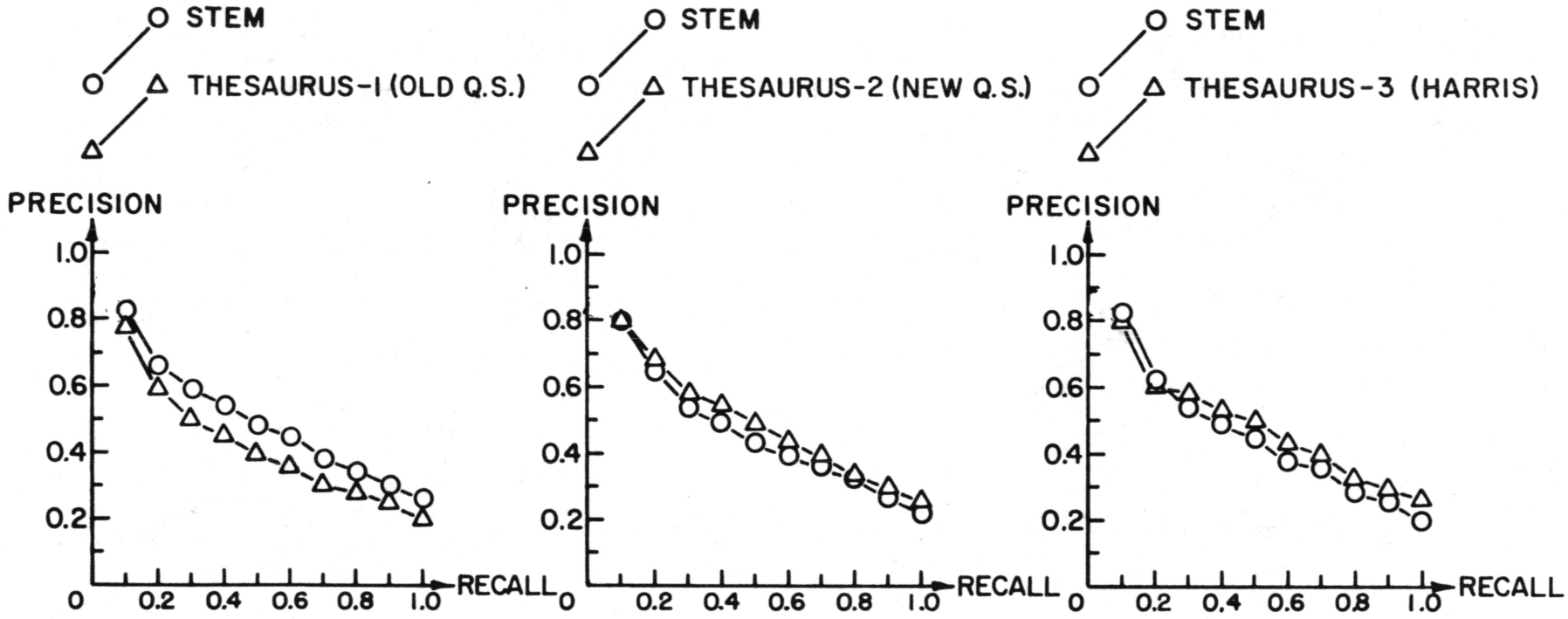
Using all comparisons of stem with the final versions of a given thesaurus, the data in Fig. 13 shows how the individual requests favor each dictionary using the normalized measures as an indicator of merit. Between 54% and 82% of the requests favor the thesaurus, with IRE-3 showing the clearest advantage for the thesaurus, ADI next, and Cran-1 the least advantage; this agrees with the precision-recall graphs.

Fig. 14 gives a further performance comparison using the average rank position of the first, second and last ranked relevant documents, to simulate high precision and high recall needs. Unexpectedly, the Cran-1 and ADI comparisons show the thesaurus to be more effective in meeting the high precision needs than stem, whereas in IRE-3, the thesaurus worsens high precision performance where one relevant only is required. However, high recall needs are seen to be dramatically



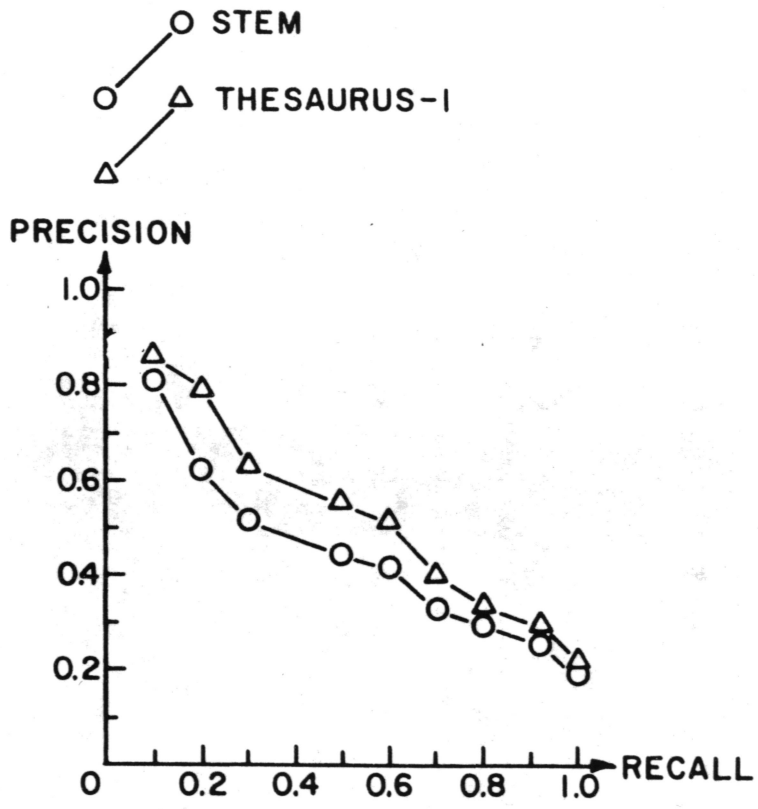
Performance Curves Comparing Stem and Thesaurus Dictionaries on the IRE-3 Collection

Fig. 8



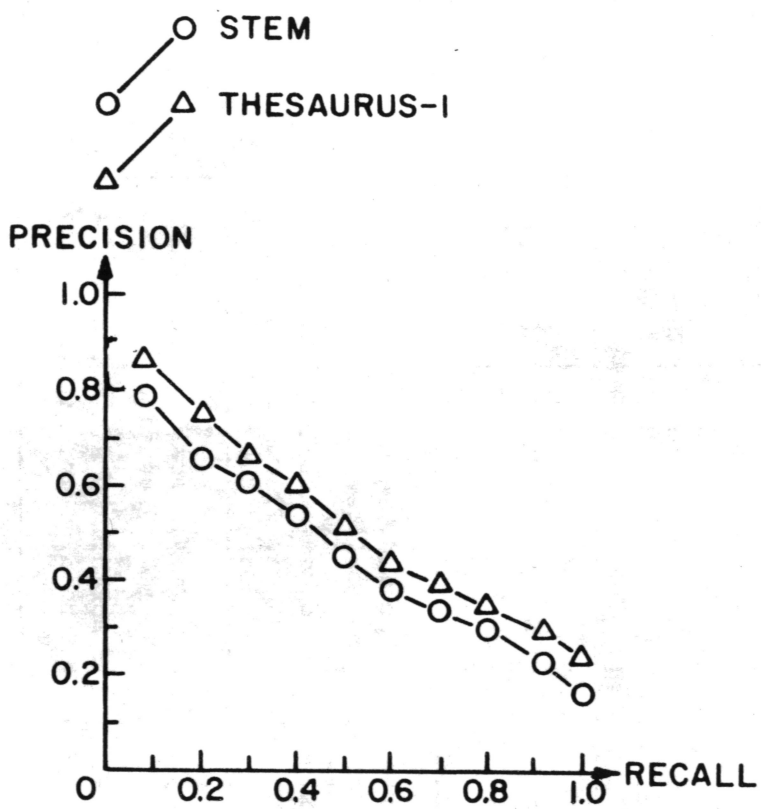
Performance Curves Comparing Stem and Thesaurus Dictionaries on the Cran-1 Collection

Fig. 9



Comparison 18

ADI, Abstracts, 35 Requests



Comparison 16

ADI, Full Text, 35 Requests

Performance Curves Comparing Stem and Thesaurus Dictionaries on the ADI Abstracts and Test Collection

Fig. 12

Collection	Input and Thesaurus Type	Evaluation Measure Used To Determine Merit	Number and Percentage* of Individual Requests					
			Thesaurus Superior		Stem Superior		Both Equal	
IRE-3 34 Requests	Abstract (Thesaurus-3)	Normed. Recall Normed. Precision	28 26	82.4% 76.5%	6 8	17.6% 23.5%	0 0	
CRAN-1 42 Requests	Abstract (Thesaurus-3) Indexing (Thesaurus-2)	Normed. Recall Normed. Precision Normed. Recall Normed. Precision	24 22 21 25	61.5% 56.4% 53.8% 64.1%	15 17 18 14	48.5% 43.6% 46.2% 35.9%	3 3 3 3	
ADI 35 Requests	Text (Thesaurus-1) Abstract (Thesaurus-1)	Normed. Recall Normed. Precision Normed. Recall Normed. Precision	22 23 20 24	64.7% 67.6% 62.5% 70.6%	12 11 12 10	35.3% 32.4% 37.5% 29.4%	1 1 3 1	

*Percentages do not include cases where dictionaries have equal merit

Comparisons of Individual Request Merit giving the Number of Requests Favoring Stem and Thesaurus for Five Runs in Three Collections, according to Merit Assigned by Normalized Recall and Precision

Fig. 13

Collection and Input	Dictionary	Average Rank of Relevant		
		First	Second	Last
IRE-3	Stem	4.4	11.6	334.0
Abstracts	Thesaurus-3	5.2	11.2	251.8
CRAN-1	Stem	7.8	13.0	72.0
Abstracts	Thesaurus-3	4.5	9.7	65.5
CRAN-1	Stem	6.4	14.8	53.3
Indexing	Thesaurus-2	4.6	12.2	47.1
ADI	Stem	8.3	12.6	36.7
Text	Thesaurus-1	5.0	10.7	33.7
ADI	Stem	8.1	15.7	39.2
Abstract	Thesaurus-1	4.3	12.2	34.5

Comparison of Stem and Thesaurus Dictionaries Using Average Rank Positions of the First, Second and Last Ranked Relevant Documents, on Five Runs in Three Collections.

Fig. 14

improved by the IRE-3 thesaurus, and hardly significantly improved over stem on Cran-1 and ADI.

Data on individual request preferences based on this average rank evaluation is given in Fig. 15. It is noteworthy that the rank position of the first relevant is unchanged by the use of a thesaurus in over one quarter of the requests. This is most strongly seen in the IRE-3 result, which shows that the drop in average rank of the first relevant with the thesaurus is caused by only very few requests being inferior to stem. The only small reversal of merit in Fig. 15 is the Cran-1 indexing result using the average rank of the last relevant, where it is seen that on an individual request basis, stem has a slight edge over thesaurus.

The use of mean rank position as in Fig. 14, is not very well suited to some of the data presented. For example, the median rank position of the first relevant document is nearly always one, so additional data on the rank position of the first relevant is given in Fig. 16. Here it may be seen that the thesaurus dictionaries all produce results for which two to six more of the requests have their first relevant in rank positions one or two; in the Cran-1 and ADI collections, the number of requests having the first relevant ranked later than ten is also reduced by the thesaurus.

The results in Figs. 6 and 7 which were based on matching functions other than cosine numeric are not presented in the form of complete precision recall graphs, but a simplified table giving the merit at three positions on the precision-recall curves appears in Fig. 17. In general, the merit is the same as that seen for the normalized measures: the cases where stem performs better than the thesaurus are of interest

Collection, Input and Thesaurus Type	Evaluation, Based on Average Rank, Fig. 14	Number and Percentage of Individual Requests					
		Thesaurus Superior		Stem Superior		Both Equal	
IRE-3, Abstract (Thesaurus-3), 34 Requests	First Rel.	5	14.7%	5	14.7%	24	70.6%
	Second Rel.	11	32.4%	4	11.8%	19	55.9%
	Last Rel.	27	79.4%	7	20.6%	0	-
CRAN-1, Abstract (Thesaurus-3), 42 Requests	First Rel.	12	28.6%	9	21.4%	11	26.2%
	Second Rel.*	14	34.1%	12	29.3%	15	36.6%
	Last Rel.	24	57.1%	15	35.7%	7	16.7%
CRAN-1, Indexing, (Thesaurus-2), 42 Requests	First Rel.	15	35.7%	7	16.7%	20	47.6%
	Second Rel.*	21	51.2%	7	17.1%	13	31.7%
	Last Rel.	18	42.9%	21	50.0%	3	7.1%
ADI, Text, (Thesaurus-1), 35 Requests	First Rel.	17	48.6%	6	17.1%	12	34.3%
	Second Rel.†	16	51.6%	8	25.8%	7	22.6%
	Last Rel.	20	57.1%	13	37.1%	2	5.7%
ADI, Abstract, (Thesaurus-1), 35 Requests	First Rel.	19	54.3%	7	20.0%	9	25.7%
	Second Rel.†	16	51.6%	11	35.5%	4	12.9%
	Last Rel.	20	57.1%	13	37.1%	2	5.7%

* In the Cran-1 Collection, 1 request has no second relevant, so results are based on 41 requests.

† In the ADI Collection, 4 requests have no second relevant, so results are based on 31 requests.

Comparison of Individual Request Merit Giving the Numbers of Requests Favoring Stem and Thesaurus on Five Runs in Three Collections, according to Merit Assigned by the Average Ranks given in Figure 14

Fig. 15

Collection and Input	Dictionary	Number of Requests with Rank of First Relevant=		
		1-2	3-10	>10
IRE-3 Abstracts	Stem	27	6	1
	Thesaurus-3	30	2	2
CRAN-1 Abstracts	Stem	24	12	6
	Thesaurus-3	26	11	5
CRAN-1 Indexing	Stem	24	10	8
	Thesaurus-2	27	10	5
ADI Text	Stem	18	8	9
	Thesaurus-1	21	9	5
ADI Abstract	Stem	16	9	10
	Thesaurus-1	22	9	4

[e.g., in IRE-3 Abstracts Stem, for 27 of the requests, the first ranked relevant document occupies rank position 1 or 2; for 6 of the requests, it occupies ranks 3 to 10, and for one request it occupies a rank larger than 10]

Comparison of Individual Request Merit Giving the Numbers of Requests Achieving Three Ranges of Rank Position to the Best Ranked Relevant, for Stem and Thesaurus Dictionaries and Five Runs in Three Collections

Fig. 16

No.	Collection	Input, Matching Function and Thesaurus Type	Stem(s) versus Thesaurus(t) Precision at Recall .2	.5	.8	Normalized Rec.	Normalized Pre.
3	IRE-3 34 Requests	Title, Cosine Numeric, Thes-2	T	T	[T]	T	T
4		Title, Cosine Numeric, Thes-3	T	[T]	[T]	T	[T]
5		Abstract, Cosine Logical, Thes-3	T	T	T	T	T
6		Abstract, Overlap Logical, Thes-2	S	S	T	T	S
7		Abstract, Overlap Logical, Thes-3	T	T	T	T	T
11	Cran-1 42 Requests	Title, Cosine Numeric, Thes-3	S	S	[T]	T	T
14		Abstract, Cosine Logical, Thes-3	T	[T]	[T]	T	[T]
15		Abstract, Overlap Logical, Thes-3	[T]	[T]	T	T	T
17		Text, Cosine Numeric, Thes-SA1	[S]	[S]	S	S	S
19		Abstract, Cosine Numeric, Thes-SA1	S	S	S	S	S

The difference of one dictionary merit over the other is smaller than 0.05 when a letter alone appears, and larger than 0.05 when letters are circled.

Table Summarizing 18 Precision Versus Recall Plots not Presented, Comparing Stem and Thesaurus Dictionaries on Three Collections.

Fig. 17

No.	Collection	Input, Matching Function and Thesaurus Type	Stem(s) versus Thesaurus(t) Precision at Recall .2	.5	.8	Normalized Rec.	Normalized Pre.
20		Title, Cosine Numeric, Thes-1	[T]	[T]	[T]	[T]	[T]
21		Title, Cosine Numeric, Thes-SA1	S	T	T	T	T
22	ADI	Text, Cosine Logical, Thes-1	S	[S]	S	T	S
23	35	Text, Cosine Logical, Thes-SA1	[S]	[S]	S	[S]	[S]
24	Requests	Abstract, Cosine Logical, Thes-1	T	T	T	T	[T]
25	(contd)	Text, Overlap Logical, Thes-1	[S]	[S]	[S]	S	[S]
26		Text, Overlap Logical, Thes-SA1	[S]	[S]	[S]	[S]	[S]
27		Abstract, Overlap Logical, Thes-1	T	S	T	T	T

The difference of one dictionary merit over the other is smaller than 0.05 when a letter alone appears, and larger than 0.05 when letters are circled.

Table Summarizing 18 Precision Versus Recall Plots not Presented, Comparing Stem and Thesaurus Dictionaries on Three Collections

Fig. 17

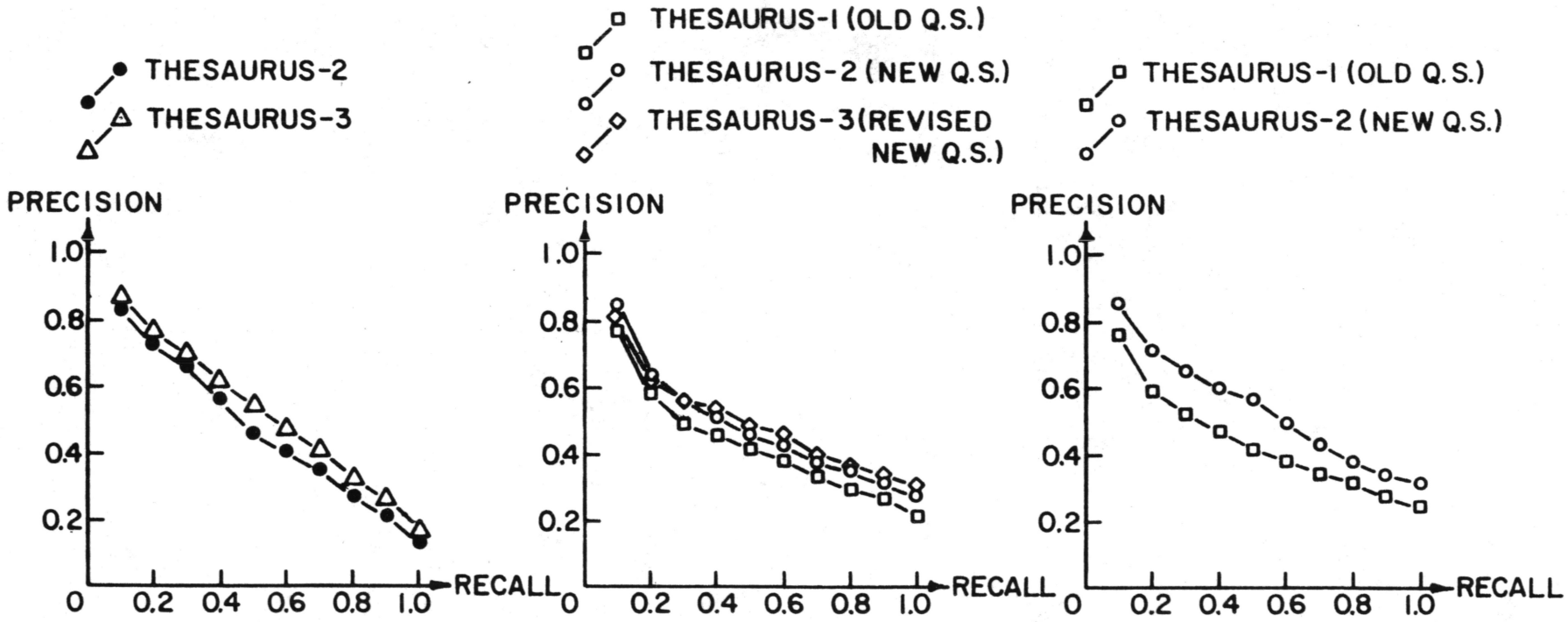
(continued)

only from a performance analysis viewpoint, since the combinations of document lengths (e.g. titles), overlap correlation and logical vectors are known to be inferior to the regular abstracts cosine numeric results.

Fig. 18 presents data already given in Figs. 8, 9, and 11, but here the performance of different versions of essentially the same dictionary may be compared, the latter version always producing some improvement.

B) Phrase and Hierarchy Dictionaries

Since both phrase and hierarchy dictionaries are based on the grouping made within a given thesaurus, performance comparisons will be made between the thesaurus alone on the one hand, and the thesaurus used with either phrases or hierarchy on the other. Using the normalized evaluation measures, four comparisons involving phrases are given in Fig. 19, and five comparisons with hierarchy appear in Fig. 20. For the phrase results in Fig. 19, phrase concept numbers are added to the requests and documents and given a weight of 1.0, equal to the weight of the original concepts in requests and documents. Phrases perform better than thesaurus on the IRE-3 collection, and on ADI, a small improvement for phrases is evident. With the Cran-1 collection, phrases perform a little worse than the thesaurus. The hierarchy results in Fig. 20 are based only on one particular series of relations searched, in which both requests and documents are expanded by means of the hierarchy, and new concepts added are given a weight of 1.0, equal to the weight of original concepts in the requests and documents. Fig. 20 shows that use of the "Sons", "Brothers" and "Cross References" relations in the hierarchy results in a near equivalent, or worse performance than the



Performance Curves Comparing Different Thesaurus Dictionaries on Two Collections

Fig. 18

Collection	Input and Type of Thesaurus	Evaluation Measure	Thesaurus-3 Dictionary	Thesaurus-3 with Phrases, Weight 1.0
IRE-3 34 Requests	Abstract (Thesaurus-2)	Normed. Recall Normed. Precision	.9191 .7072	.9282 .7252
	Abstract (Thesaurus-3)	Normed. Recall Normed. Precision	.9268 .7382	.9326 .7529
CRAN-1 42 Requests	Abstract (Thesaurus-3)	Normed. Recall Normed. Precision	.8837 .6952	.8791 .6873
ADI 35 Requests	Text (Thesaurus-1)	Normed. Recall Normed. Precision	.8206 .6273	.8224 .6336

Performance Results Comparing Thesaurus Without and with Phrases for Four Results Using Cosine Numeric on Three Collections, and Normalized Recall and Precision

Fig. 19

Collection	Input	Evaluation Measure	Hierarchy Search	Hierarchy-3 Dictionary	Thesaurus-3 Dictionary
IRE-3 34 Requests	Abstract	Normed. Recall Normed. Precision	Parents	.9437 .7600	.9268 .7382
	Abstract	Normed. Recall Normed. Precision	Sons	.9269 .7129	.9268 .7382
	Abstract	Normed. Recall Normed. Precision	Brothers	.9220 .6972	.9268 .7382
	Abstract	Normed. Recall Normed. Precision	Cross Refs.	.9229 .7134	.9268 .7382
	Abstract	Normed. Recall Normed. Precision	All Relations, (Parents, Sons, Brothers and Cross Refs.)	.9446 .7506	.9268 .7382

All Hierarchy runs use a weight of 1.0, and expand both requests and documents

Performance Results Comparing Thesaurus and Hierarchy Dictionaries for Five Results using Cosine Numeric on the IRE-3 Collection, and Normalized Recall and Precision

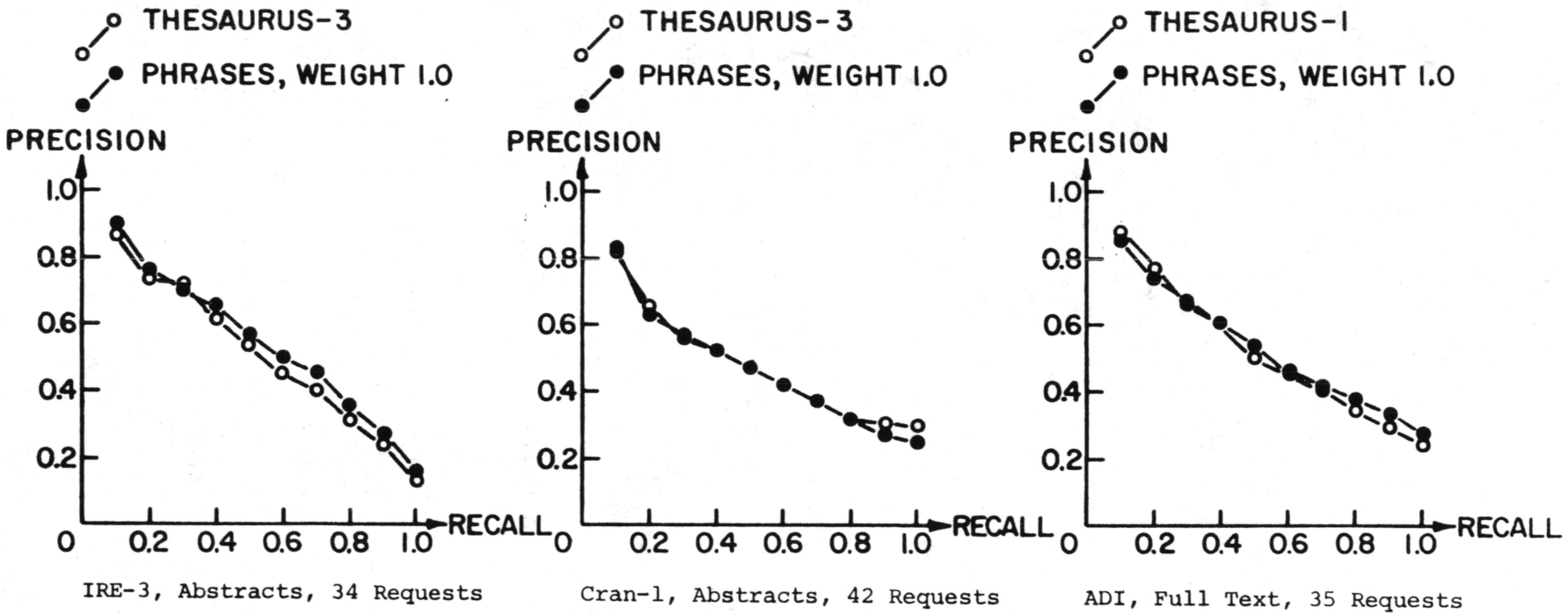
Fig. 20

thesaurus alone. The use of "Parents" and "All Relations" does, however, give some performance increase over the thesaurus. Previous tests on the same hierarchy using the IRE-2 collection with 17 requests showed the hierarchy to be always inferior to the thesaurus, (two of the results appear in [16]).

Precision versus recall graphs are to be seen in Figs. 21, 22, and 23. The phrase results in Fig. 21 agree with the results based on the normalized measures; in the case of IRE-3 and ADI, the phrase superiority is very small indeed. The hierarchy results in Figs. 22 and 23 reveal that the "Parents" and "All Relations" results are superior to the thesaurus over portions of the curve only.

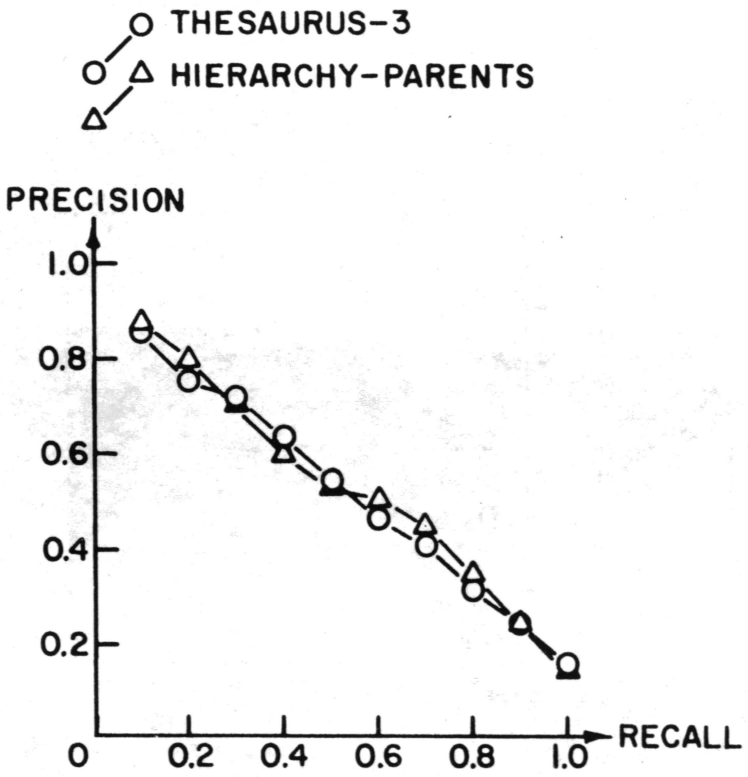
Individual request data based on the normalized measures are given for three of the phrase results and the two best hierarchy results in Fig. 24. Nearly 70% of the requests favor phrases on IRE-3, 58% favor phrases on ADI, and between 60% and 70% of the requests on Cran-1 favor the thesaurus rather than phrases. The hierarchy results show that the requests equally favor the "Parents" relation, and, unexpectedly, 62% to 67% of the requests favor the thesaurus rather than the "All Relations". This last result completely reverses the picture previously presented, and reveals that for this hierarchy option, a few requests which do very well, cause the averages to favor the hierarchy.

A comparison of merit in Fig. 25, makes use of the average rank of the first, second and last relevant documents, and the results follow the expected pattern. The average rank improvement with phrases on IRE-3 is seen to be quite small for the high precision user, however. The very large improvements in average rank of the first relevant document with both the hierarchy options is caused by only one or two requests. Fig. 26 gives

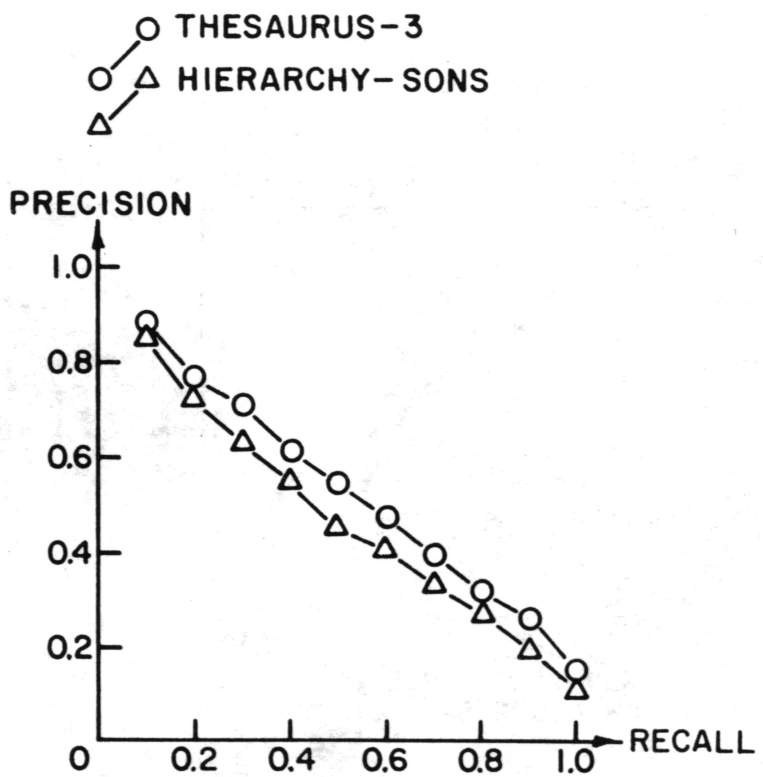


Performance Curves Comparing Thesaurus and Phrase Dictionaries on Three Collections

Fig. 21



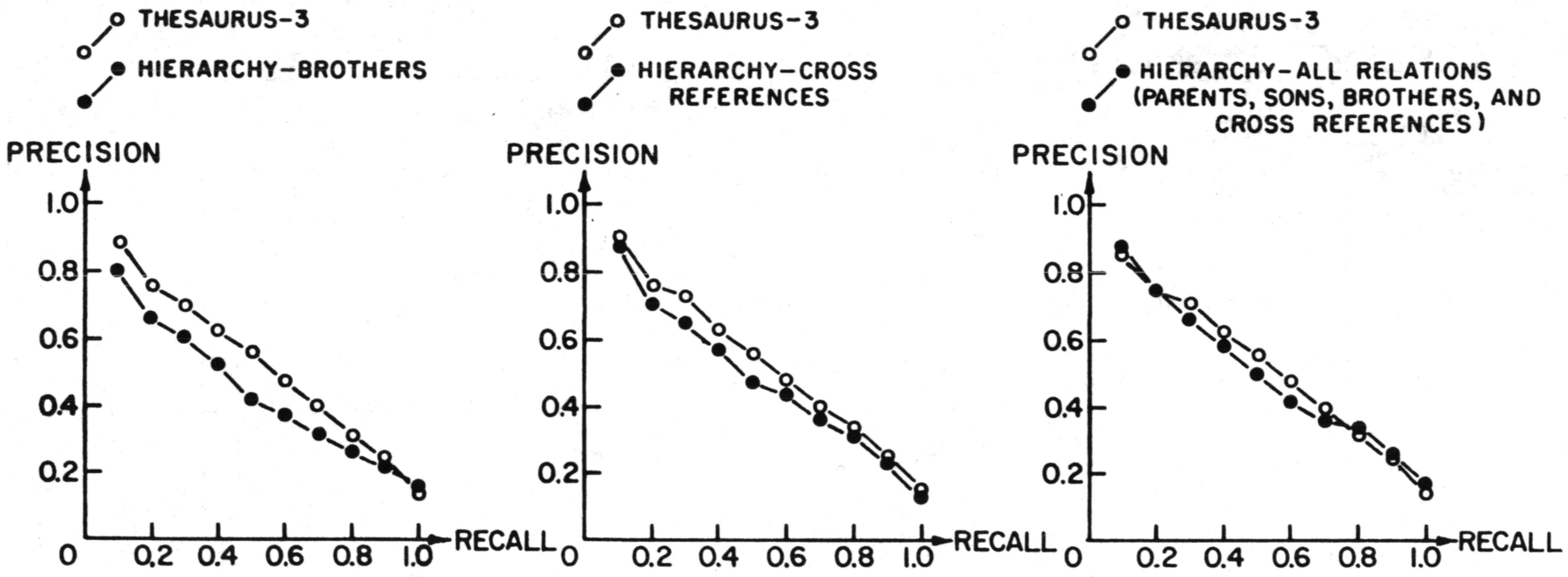
a) Hierarchy Expansion by Parents



b) Hierarchy Expansion by Sons

Performance Curves Comparing Thesaurus with Two Hierarchy Searches, IRE-3 Collection

Fig. 22



a) Hierarchy Expansion by Brothers

b) Hierarchy Expansion by Cross References

c) Hierarchy Expansion by Parents, Sons, Brothers and Cross References

Performance Curves Comparing Thesaurus with Three Hierarchy Searches, IRE-3 Collection

Fig. 23

Collection, Input and Thesaurus Type	Evaluation Measure Used To Determine Merit	Number and Percentage* of Individual Requests		
		Phrases Superior	Thesaurus Superior	Both Equal
IRE-3, Abstracts, Thesaurus-3, 34 Requests	Normed. Recall	23 67.6%	11 32.4%	0
	Normed. Precision	24 70.6%	10 29.4%	0
CRAN-1, Abstracts, Thesaurus-3, 42 Requests	Normed. Recall	11 31.4%	24 68.6%	7
	Normed. Precision	15 40.5%	22 59.5%	5
ADI, Text, Thesaurus-1, 35 Requests	Normed. Recall	16 55.2%	13 44.8%	6
	Normed. Precision	17 58.6%	12 41.4%	6
		Hierarchy Superior	Thesaurus Superior	
IRE-3 Abstracts, Thesaurus-3, 34 Requests		"Parents"		
	Normed. Recall	17 50.0%	17 50.0%	0
	Normed. Precision	17 50.0%	17 50.0%	0
		"All Relations"		
	Normed. Recall	13 38.2%	21 61.8%	0
	Normed. Precision	11 33.3%	22 66.7%	1

* Percentages do not include cases where both dictionaries have equal merit

Comparisons of Individual Request Merit Giving the Number of Requests Favoring Thesaurus and Phrases, and Thesaurus and Hierarchy, according to Merit Assigned by Normalized Recall and Precision

Fig. 24

Collection and Input	Dictionary	Average Rank of Relevant		
		First	Second	Last
IRE-3 Abstract	Thesaurus-3	5.2	11.2	251.8
	Phrases	4.6	10.4	235.2
	Hierarchy - Parents	1.6	6.2	208.3
	Hierarchy - All Relations	1.3	10.1	209.4
CRAN-1 Abstract	Thesaurus-3	4.5	9.7	66.2*
	Phrases	4.5	10.8	66.9
ADI Text	Thesaurus-1	5.0	10.7	33.7
	Phrases	5.7	9.9	33.9

* This result differs somewhat from that given in Fig. 14 because this comparison requires the use of output assigning a slightly different rank in cases of tied rank positions.

Comparison of Thesaurus with Phrases and Thesaurus with Hierarchy Using the Average Rank Positions of the First, Second and Last Ranked Relevant Documents.

Fig. 25

individual request preferences based on Fig. 25, and for every test, there exists for at least one of the three evaluation ranks, a case where the thesaurus is to be preferred to the phrases or hierarchy. Fig. 27 shows more clearly how, using results based on the average rank of the first relevant recovered (to simulate a high precision user) phrases are not superior to thesaurus in any of the results, but the hierarchy relations do give a very small improvement.

6. Summary of Results

Since the volume of data tends to obscure overall findings, the results of performance comparisons are enumerated separately. In order to facilitate reference to the thesaurus results, the 28 comparisons made are referred to by number; the normalized evaluation results may be found in Figs. 5, 6, 7, and 10, and the precision versus recall results in Figs. 8, 9, 10, 11, 12, and 17. The thesaurus results may be summarized as follows:

1. The best thesaurus dictionaries give a performance superior to the stem dictionary on the average if other system parameters are set to their optimum (using abstracts or text, together with the cosine numeric matching function). This is seen in five cases (comparisons 2, 10, 13, 16, and 18); the superiority of thesaurus is least marked in the Cran-1 collection.

2. The Cran-1 collection is unique in that the suffix 's' dictionary performs a little better than stem; a comparison of thesaurus with suffix 's' shows the suffix 's' to be a little superior at the high precision end of the curve (comparison 28, Fig. 10).

Collection, Input and Type, Thesaurus	Evaluation Based on Average Rank	Number and Percentage of Individual Requests		
		Phrases Superior	Thesaurus Superior	Both Equal
IRE-3, Abstract, Thesaurus-3, 34 Requests	First Rel.	4 11.8%	4 11.8%	26 76.5%
	Second Rel.	4 11.8%	8 23.5%	22 64.7%
	Last Rel.	16 47.1%	14 41.2%	4 11.8%
CRAN-1, Abstract, Thesaurus-3, 42 Requests	First Rel.	10 23.8%	12 28.6%	20 47.6%
	Second Rel.*	10 24.4%	15 36.6%	16 39.0%
	Last Rel.	13 31.0%	16 38.1%	13 31.0%
ADI, Text, Thesaurus-1 35 Requests	First Rel.	3 8.6%	7 20.0%	25 71.4%
	Second Rel.†	12 38.7%	4 12.9%	15 48.4%
	Last Rel.	12 34.3%	13 37.1%	10 28.6%
		Hierarchy Superior	Thesaurus Superior	Both Equal
IRE-3, Abstract, Thesaurus-3, 34 Requests		"Parents"		
	First Rel.	5 14.7%	2 5.9%	27 79.4%
	Second Rel.	7 20.6%	7 20.6%	20 58.8%
	Last Rel.	15 44.1%	19 55.9%	0 -
		"All Relations"		
	First Rel.	6 17.6%	5 14.7%	23 67.6%
	Second Rel.	6 17.6%	13 38.2%	15 44.1%
	Last Rel.	13 38.2%	21 61.8%	0 -

* In the Cran-1 Collection, 1 request has no second relevant, so results are based on 41 requests.

† In the ADI collection, 4 requests have no second relevant so results are based on 31 requests.

Comparison of Individual Request Merit giving the Number of Requests Favoring Thesaurus and Phrases, and Thesaurus and Hierarchy, according to Merit Assigned by the Average Ranks given in Fig. 25

Fig. 26

Collection and Input	Dictionary	Number of Requests with Rank of First Relevant =		
		1-2	3-10	≥ 11
IRE-3 Abstracts	Thesaurus-3	30	2	2
	Phrases	30	2	2
	Hierarchy - Parents	30	3	1
	Hierarchy - All Relations	32	2	-
CRAN-1 Abstract	Thesaurus-3*	27	10	5
	Phrases	26	11	5
ADI Text	Thesaurus-1	21	9	5
	Phrases	20	9	6

* This results differs slightly from that given in Fig. 16 because the present comparison requires the use of output that assigns a slightly different rank in cases of tied rank positions.

Comparison of Individual Request Merit giving the Number of Requests Achieving Three Ranges of Rank Position to the Best Ranked Relevant, comparing Thesaurus with Phrases, and Thesaurus with Hierarchy

Fig. 27

CRAN-1 Collection

Request	Stem Dictionary				Thesaurus-3 Dictionary			
	Rank	Document		Corr.	Rank	Document		Corr.
Q79 3 rel.	132	302	Relevant	.0000	32	436	Relevant	.1114
	132	436	Relevant	.0000	33	437	Relevant	.1108
	132	437	Relevant	.0000	145	302	Relevant	.0000
	Normed. Recall .3401 Normed. Precision .0874				Normed. Recall .6548 Normed. Precision .2797			
Q225 6 rel.	25	655	Relevant	.2139	5	07F	Relevant	.4327
	40	569	Relevant	.1892	8	569	Relevant	.3813
	42	07F	Relevant	.1771	11	655	Relevant	.3750
	58	687	Relevant	.1283	54	572	Relevant	.2306
	66	572	Relevant	.1051	93	687	Relevant	.1404
	169	07G	Relevant	.0000	167	07G	Relevant	.0366
	Normed. Recall .6744 Normed. Precision .3059				Normed. Recall .7277 Normed. Precision .4769			

Fig. 28

CRAN-1 Collection

Request	Stem Dictionary				Thesaurus-3 Dictionary			
	Rank	Document		Corr.	Rank	Document		Corr.
Q167 4 rel.	3	916	Relevant	.2564	1	728	Non-relevant	.4045
	13	921	Relevant	.2055	15	919	Relevant	.2302
	19	919	Relevant	.1784	18	916	Relevant	.2107
	41	920	Relevant	.1267	21	921	Relevant	.2052
	>41	728	Non-relevant	.0289	62	920	Relevant	.1375
	Normed. Recall .9158 Normed. Precision .6028				Normed. Recall .8648 Normed. Precision .4667			
Q323 5 rel.	1	34A	Relevant	.2146	1	316	Non-relevant	.4173
	13	33H	Relevant	.1304	13	34A	Relevant	.2494
	43	33I	Relevant	.0597	29	33H	Relevant	.1739
	65	34+	Relevant	.0393	72	33I	Relevant	.1043
	147	879	Relevant	.0000	106	34+	Relevant	.0685
	147	316	Non-relevant	.0000	177	879	Relevant	.0000
	Normed. Recall .7395 Normed. Precision .5057				Normed. Recall .6082 Normed. Precision .2952			

Retrieval Results of Four Individual Requests Comparing Stem with Thesaurus

Fig. 28

(continued)

3. The initial versions of a thesaurus, and dictionaries without the construction rules are inferior to revisions and versions made using the rules; and in two out of seven comparisons, the performance of the initial thesaurus versions is not as good as the stem process (comparisons 8 and 12 worse than stem; comparisons 1, 3, 6, 9 and 13 superior to stem).

4. The thesaurus superiority is not always preserved when less than optimal document length and matching function parameters are used; thus, in twelve comparisons, three are inferior to stem (comparisons 22, 25, and 26 inferior; comparisons 4, 5, 7, 11, 14, 15, 20, 24, and 27 superior).

5. For users needing high precision with only one or two relevant documents, the thesaurus is little better than stem on IRE-3, but in Cran-1 and ADI, a larger superiority for the thesaurus is evident (see Figs. 14, 15, and 16).

6. For users with a very high recall need, IRE-3 produces a good improvement for the thesaurus over stem, but in Cran-1 and ADI only a very small gain is seen, using the average rank of the last relevant document as a measure (Figs. 14 and 15).

7. The thesaurus-SA1 on ADI, made by the semi-automatic rules, does not provide a good performance. It is in all cases inferior to the ADI regular thesaurus-1, and in four of five comparisons it is also inferior to stem (Comparisons 17, 19, 23 and 26 inferior; comparison 21 superior).

Results comparing the thesaurus with the addition of phrases are as follows:

1. Phrase dictionaries give a superior performance compared with thesaurus alone by a very small amount only on IRE-3 and ADI, and on Cran-1 the thesaurus alone gives a slightly better result (Figs. 19 and 21).

2. For a high precision need, only IRE-3 produces some advantage to phrases (Figs. 21 and 25); however, this is based on a small superiority for one or two requests only, and is not considered significant (Figs. 26 and 27).

3. For a high recall need, use of the average rank of the last relevant shows the phrases to be useful on IRE-3 only (Fig. 25) by a small margin on an individual request basis (Fig. 26).

Results comparing the thesaurus with the addition of various hierarchy relations on the IRE-3 collection produce the following conclusions:

1. Thesaurus alone is always superior to hierarchy on three of the relations tested, and on two others ("parents and "all" relations), the hierarchy gives a small advantage over portions of the precision recall curve (Figs. 22, 23). On an individual request basis (Fig. 24), the thesaurus is equal to "parents", and superior to "all" relations; the hierarchy is thus not to be preferred.

2. For a high precision need, Fig. 25 suggests that some advantage accrues, but Fig. 27 shows that its success is limited to one or two requests that do badly with the thesaurus alone.

3. For a high recall need, Fig. 25 shows that the hierarchy performs well, but Fig. 26 reveals again that it achieves only a few dramatically good results with a poorer average high recall performance for individual requests than thesaurus.

7. Performance Analyses

The first task of the analysis is to explain the mechanism which causes an improvement in retrieval performance using the thesaurus and

also to consider cases where the thesaurus worsens performance. Retrieval results from four of the Cran-1 requests are given in Fig. 28, using the suffix 's' dictionary and the thesaurus-3 dictionary. Requests Q79 and Q225 have an overall superiority on thesaurus, and requests Q167 and Q323 prefer suffix 's'. The thesaurus improvement for documents 436 and 437 in Q79 is reflected in the size of the correlation coefficient; this is due to some thesaurus produced matches between request and document, when suffix 's' produced no matches at all. In Q225, document 07F is improved in rank by 37 places using the thesaurus, because the request contains a hyphenated phrase "Boundary-layer" which was matched by the thesaurus with the occurrence of the component words occurring separately in the document. This is an instance where the suffix 's' or stem dictionary could cope with the problem of hyphens were disregarded. The superiority of thesaurus over stem would then be reduced from 0.0193 to 0.0053 in normalized recall, and 0.0248 to 0.0141 normalized precision (Cran-1) when hyphens are removed.

A quite different way in which the thesaurus improves performance is illustrated by document 655 in Q225; this item increases by 17 positions in rank. Both suffix 's' and thesaurus provide three matches between request and document concepts, but the match with the concept "Boundary-layer" receives a weight of 5 with the thesaurus and only 1 with suffix 's'. The numeric vector weighting produced by the thesaurus thus proves effective in this case; and the thesaurus with weights in fact acts as a precision device. In fact, document 569 and 572 are improved by the thesaurus for the same reason, thus showing that the thesaurus proves

superior not always through the introduction of additional matching concepts.

Cases of the superiority of suffix 's' over thesaurus are also shown in Fig. 28, Q167 and Q323. For example, relevant document 916 matches with five request concepts for both suffix 's' and thesaurus; but since the thesaurus process fails to match with any additional request concepts, and also provides no increase in the weight of any of the matching concepts, document 916 is relegated in rank by non-relevant documents such as 728. In the case of 728, which matches one concept on suffix 's' only, the thesaurus provides additional matching concepts; also since 728 is a short document, it produces a high cosine correlation coefficient and receives the first rank position. In Q323, non-relevant document 316 is matched by four concepts with the thesaurus, and although the thesaurus establishes two additional matches with relevant document 34A, this is not sufficient to prevent non-relevant documents from occupying the top rank positions.

These examples from the Cran-1 collection lead to the question of whether the lessened superiority of thesaurus over stem compared with IRE-3 and ADI is due to a poor thesaurus dictionary or to something in the Cran-1 test environment. Evidence strongly points to the latter reason. Cran-1 has real user relevance decisions that, on inspection, provide a severe test environment and use relevance decisions that sometimes bear little relation to the stated request. The superiority of suffix 's' over the stem dictionary is not found on IRE-3 and ADI; in Section V, the reason for this is stated to be the terminology employed

in aerodynamics. This does not rule out the factor of quality of the thesaurus, but since the thesaurus-3 was constructed with the thesaurus rules by the same person who made the IRE thesaurus-3 and ADI thesaurus-1, it does not seem likely that the Cran thesaurus-3 is really bad. A further reason for accepting the Cran-1 environment as being responsible lies in the findings of the Cranfield Project [17] in which the "quasi-synonym" index language was found to be inferior to the "word form" (i.e. stem) language. The quasi-synonym list was rearranged in certain respects to become the SMART Cran thesaurus-1, and this dictionary does perform worse than stem for both abstracts and indexing. The reason for this result offered by the Cranfield Project was that imprecise terminology is not helped by a dictionary grouping of the thesaurus type since any help given to the poorly matched relevant documents is offset by an increased number of retrieved non-relevant documents. This conclusion needs modification in two ways, in the light of the SMART results. First, terminology alone is unlikely to be the only explanation, since the ADI collection on documentation is believed to use terminology at least as imprecise as aerodynamics, and the thesaurus-1 does provide some advantage for that collection. Second, some amount of grouping of aerodynamics terms does give a slight performance increase, since the thesaurus-3 dictionary gives better results than stem or suffix 's' over some parts of the curve. Another point on which the Cranfield conclusion may not apply to SMART is in the use by SMART of a weighting scheme of the type not tested at Cranfield. In fact, without the weighting scheme in use, the ADI <u>text</u> result does follow the Cranfield

conclusion (comparisons 22 and 25), although ADI abstracts do not. It is suggested therefore that a thesaurus with weights does have some additional power, probably due to the precision device effect that has been illustrated.

Two examples from the other collections are given in Fig. 29. The ADI request QB10 has a worse than random normalized recall using the stem dictionary, and the large improvements achieved by thesaurus are due mainly to the new synonym connection between "computerization" and "computer" (not confounded by stem), the dropping of the word "system" by making it a restricted word in the thesaurus, and the very large increase in weight of important concepts such as "chemistry", due to the synonym groupings. If a small amount of human intervention in the weighting scheme were permitted, a simple increase of three in the weight of the one vital request concept "chemistry" would result in a thesaurus result of ranks 1, 2, 3, and 5 for the four relevant. The IRE-3 example shows cases of relevant documents considerably worsened in rank by the thesaurus. In the case of documents 200 and 382, for example, the thesaurus provides no increase in weight to any of the concepts that matched on stem, and furnishes only one additional matching concept. Also, the word "method" is dropped from the thesaurus, an apparently sensible decision, but this highly weighted term matched the request using the stem process, thus helping the result.

These individual examples show that a considerable amount of variation in individual requests is obscured by the use of averages alone. This suggests that some method of making an accurate pre-search dictionary choice would produce good results; attempts to come up with such a method have, however, not succeeded so far.

Collection and Request	Stem Dictionary			Thesaurus Dictionary		
	Rank	Relevant Document	Corr.	Rank	Relevant Document	Corr.
ADI QB10 4 rel.	33	09	.1512	1	09	.4100
	48	70	.0977	7	48	.3062
	58	69	.0769	26	70	.1825
	67	48	.0599	32	69	.1268
	Normed. Recall .3718			Normed. Recall .8205		
	Normed. Precision .1336			Normed. Precision .6180		
IRE-3 Q015 6 rel.	9	200	.3539	106	200	.2448
	71	382	.2503	153	106	.2044
	212	106	.1630	189	382	.1815
	309	71H	.1210	283	71H	.1390
	498	85A	.0563	301	85A	.1315
	691	72+	.0000	669	72+	.0000
	Normed. Recall .6191			Normed. Recall .6382		
	Normed. Precision .2889			Normed. Precision .2141		

Retrieval Results for Two Individual Requests Comparing Stem with Thesaurus

Fig. 29

Evaluation of the semi-automatic "Hastie" thesaurus-SAl on ADI must await the testing of a further version of this thesaurus. However, the tentative conclusions are that this method is not workable in practice, owing to the difficulty of generating suitable property questions, and the need to re-sort the resulting groups using frequency information. Some further developments may provide solutions to these problems.

Examination of individual requests using the phrases shows that no dramatic performance changes take place, and in general, the phrases do not give a significant advantage even for the IRE-3 collection. Part of the reason for this is the small number of phrases included in the dictionaries. Also, use of phrases to replace the occurrences of the individual component concepts would probably alter the request and document vectors by a greater amount than the present procedure of simply adding phrase concepts; performance changes (either better or worse) would then be more clearly seen.

Results using the hierarchy show it to be very effective for only a few individual requests. An examination of all requests immediately shows that the 17 staff prepared requests behave differently from the 17 non-staff prepared ones, and Fig. 30 shows that there is a tendency for hierarchy to be more effective on the non-staff requests than the staff ones. It was seen in Section I that the staff requests have a much better performance than the non-staff requests, therefore there is less room for improvement with hierarchy for these requests, and the extra hierarchy identifiers only serve to increase the match with non-relevant documents. The non-staff requests have exhibited a poor performance with thesaurus, and thus leave room for improvement by additional dictionary grouping (which is what the hierarchy does).

Collection	Request Type	Evaluation Measure used to Determine Merit	Number and Percentage* of Individual Requests		
			Thesaurus-3 Superior	Hierarchy-3 Superior	Both Equal
IRE-3				"Parents"	
	17 Staff-Prepared Requests	Normed. Recall	10 58.8%	7 41.2%	0
		Normed. Precision	14 82.4%	3 17.6%	0
	17 Non-Staff Prepared Requests	Normed. Recall	7 41.2%	10 58.8%	0
		Normed. Precision	8 47.1%	9 52.9%	0
				"All Relations"	
	17 Staff-Prepared Requests	Normed. Recall	12 70.6%	5 29.4%	0
		Normed. Precision	14 82.4%	3 17.6%	0
	17 Non-Staff Prepared Requests	Normed. Recall	8 47.1%	9 52.9%	0
		Normed. Precision	8 50.0%	8 50.0%	1

* Percentages do not include cases where both dictionaries have equal merit.

Comparison of Individual Request Merit Giving the Number of Requests Favoring Thesaurus and Hierarchy Contrasting the Staff and Non-Staff Prepared Requests on IRE, According to Merit Assigned by Normalized Recall and Precision

Fig. 30

Examples of the improvement given to two of the non-staff requests by the hierarchy are given in Fig. 31. Request Q006 asks for documents about information retrieval using computers, and concept 26 "retrieval" is linked in the hierarchy to "parent" concept 200 "data-processing", "data handling", etc. All six relevant documents also contain concept 200 as a result of the hierarchy expansion; one document did not originally contain concept 26, and so obtained concept 200 from "sums" other than concept 26; the other documents achieved high weights on concept 200 through a similar connection. Thus, concept 200 is in the main responsible for the sharp improvement in performance, mainly through the mechanism of increasing the weight given to the notion vital to the request.

Request Q015 has six concepts in the request when the thesaurus is used, and this is expanded to twenty-six when the hierarchy "all" relation is in use. Document 200 has a greatly improved rank on hierarchy, because all but two of the additional request concepts added by hierarchy are matched, thus giving a total match of 23 out of 26 on hierarchy, although 5 out of 6 matches were achieved by thesaurus. In general, it is clearly unusual for a document to match with nearly all the hierarchy expansions in a given request, and the case of document 200 may be a special one. Documents 106 and 382 both exhibit cases of hierarchy acting as a recall device, since request concepts 383 ("Transcendental") and 618 ("Function") do not match with the thesaurus, but do match with hierarchy through "brothers" and "cross reference" relations.

This points to the probably reason why the hierarchy as tested is not generally effective: because the use of thesaurus groups to build

Collection and Request	Thesaurus-3 Rank	Relevant Document	Corr.	Hierarchy-3, "Parents" Rank	Relevant Document	Corr.
IRE-3 Q006 5 rel.	24	221	.2911	1	221	.4233
	58	080	.2275	10	080	.3095
	65	126	.2210	16	126	.2506
	86	28B	.2032	34	079	.2112
	397	079	.0550	37	28B	.2108
	Normed. Recall .8413			Normed. Recall .9786		
	Normed. Precision .4012			Normed. Precision .7394		
	Thesaurus-3			Hierarchy-3 "All Relations"		
IRE-3 Q015 6 rel.	106	200	.2448	1	106	.7853
	153	106	.2044	3	382	.7448
	189	382	.1815	14	85A	.6516
	283	71H	.1390	47	200	.5203
	301	85A	.1315	94	71H	.3916
	669	72+	.0000	492	72+	.0647
	Normed. Recall .6382			Normed. Recall .8649		
	Normed. Precision .2141			Normed. Precision .6480		

Retrieval Results of Two Individual Non-Staff Prepared Requests Comparing Thesaurus with Hierarchy

Fig. 31

a hierarchy brings in too many words, and permits combinations of individual words to be compounded that give no useful grouping for retrieval. A hierarchy based on, say, the stem dictionary might give better results, and tests of a hierarchy based on suffix 's' will be made for the Cran-1 collection, but this particular Cran-1 hierarchy (constructed at Cranfield) is very difficult to construct, and it did not perform well at Cranfield. Since hierarchies are normally based partly on phrases and partly on single words, any new work in phrase processing would provide a much more interesting environment, in which a hierarchy could be constructed and tested. The inclusion of a hierarchy within an automatic system does seem to require the user to examine portions of the hierarchy in relation to their particular search request, since the many optional uses of hierarchy, such as "parents", "sons" etc. would require some definite pre-search choice of the relation to be used.

This analysis and discussion of the phrases and hierarchy has shown that, in their present form, these two types of dictionary do not improve the thesaurus process by an amount that would justify the effort required for construction. Indeed, it might even be questioned whether the effort of constructing a thesaurus itself is worthwhile, since results such as those given in Figs. 14, 15, and 16 prove that the improvement of performance in comparison with the stem dictionary is not really large. In situations where economic considerations are all important, or time is very limited, it seems that an automatic stem dictionary will perform quite well, particularly for the high precision user. It is disappointing that the thesauruses tested do not always help the high recall user;

Collection	Input and Thesaurus Type	Evaluation Measure	Performance Differences, + = Thesaurus Superior, - = Stem Superior Cosine Numeric	 Cosine Logical
IRE-3 34 Requests	Abstract (Thesaurus-3)	Normed. Recall Normed. Precision	+ .0314 + .0636	+ .0290 + .0407
CRAN-1 42 Requests	Abstract (Thesaurus-3)	Normed. Recall Normed. Precision	+ .0193 + .0248	+ .0332 + .0559
ADI 35 Requests	Text (Thesaurus-1)	Normed. Recall Normed. Precision	+ .0427 + .0700	+ .0124 - .0156
	Text (Thesaurus-SA1)	Normed. Recall Normed. Precision	- .0005 - .0132	- .0811 - .0916
	Abstract (Thesaurus-1)	Normed. Recall Normed. Precision	+ .0415 + .0743	+ .0497 + .0602

Performance Differences Between Stem and Thesaurus Taken From Figs. 5 and 6 Using the Normalized Measures, Showing the Results for Numeric (Weighted) and Logical Vectors

Fig. 32

the examples given show that the thesaurus and the hierarchy are often successful because of the precision device effect achieved by the weighting scheme. Confirmation of the fact that this phenomenon is largely responsible for the improvements gained by thesaurus in the IRE-3 and ADI collections is obtained from the data of Fig. 32, where performance results from Figs. 5 and 6 are represented with, and without, the weighting process to show how the thesaurus offers greater improvement over stem when the weighting scheme is in use. The Cran-1 collection does not in this instance show this result, probably because of the effect on the cosine correlation of the change in weighting; alternatively, it may be explained as yet another instance of the difference in behavior between the Cran-1 and the others.

8. Further Studies Required

Since the conclusions of this section have already been stated in part 6, this final part enumerates some topic areas for further investigation that may be directly or indirectly suggested by the preceding analysis. Eleven studies are listed:

1. The effectiveness of all five of the dictionary construction rules must be established by the construction of a series of versions of a given dictionary, so that the relative importance of rules about word frequency versus rules about synonymy can be established. As a start in this direction, a second version of the ADI semi-automatic thesaurus is under test.

2. The present practice of reducing the weight of ambiguous terms (where "ambiguous" refers to terms grouped in more than one place in the thesaurus) should be evaluated.

3. The degree of overlap among thesaurus groups is at present kept very low, but one example of a dictionary with a large amount of overlap produced good performance; an investigation of this phenomenon is needed.

4. Thesaurus dictionaries using many terms in very few concepts do not necessarily perform poorly, as was originally believed. Unpublished results for a version of the ADI thesaurus-1 in which a further grouping of the concepts is made by statistical association to form approximately 170 concepts, gives a performance somewhat superior to the thesaurus-1 alone. The occasional examples of the value of IRE-3 hierarchy to individual requests shows that a broad grouping can sometimes work well. The relevance feedback results presented in Report ISR-12 show that very greatly expanded requests can often be used to improve the ranks of initially poorly ranked relevant documents. These examples point up the need to examine the grouping problem in depth.

5. An aid to improvements in thesaurus grouping might be the construction of a thesaurus by "hindsight"; that is, using information about given relevant documents in relation to their search requests; an optimum thesaurus might then be made in an attempt to discover more rules and principles.

6. An operational use of thesaurus-type dictionaries might be aided by the construction of "near" and "far" synonym thesauruses. The near synonym thesaurus would only contain very closely related words, and would always be used by the system, but the far synonym thesaurus would include groupings of many words that would be used only to permit a manual pre-search selection.

7. Any further testing of phrases of the type presently used, requires more exhaustive phrase lists. Tests could also be made in which phrases are given quite high weights; a strategy in which phrase identifiers would replace component concept identifiers would be of interest in this connection.

8. New types of phrase recognition procedures would be a better test of phrases than present methods.

9. New hierarchies based on stems or on some improved type of phrase procedures would require a large amount of construction effort; this possibility should not, however, be abandoned.

10. In thinking of the operational use of dictionaries, the design of methods for manually (or, automatically) generating an accurate pre-search choice from a selection of dictionaries should be undertaken. Also, the present practice of looking up both requests and documents in a given dictionary seems to be unnecessary, since no additional matches between requests and documents are established if only the requests are processed. Expanding requests only might produce some advantage in an operational user-interactive system, since some users would want to select and reject certain synonyms personally, and would not want to use rigid thesaurus groups at each stage. Evaluation of this suggestion is needed to determine the effect of expanding requests using only the weighting scheme and correlation functions in use.

11. The need to process foreign language documents has previously been pointed out; a German translation of the ADI "ISPRA" Thesaurus is currently tested in this connection.

References

[1] C. Cleverdon, J. Mills and M. Keen, Factors Determining the Performance of Indexing Systems, Vol. 1, Design, Aslib Cranfield Research Project, Cranfield, 1966.

[2] M. E. Lesk and G. Salton, Design Criteria for Automatic Information Systems, Information Storage and Retrieval, Report ISR-11, to the National Science Foundation, Section V, Department of Computer Science, Cornell University, June 1966.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] M. Lesk and T. Evslin, Statistical Phrase Processing, Information Storage and Retrieval, Report ISR-7, to the National Science Foundation, Section IX, Harvard Computation Laboratory, June 1964.

References (contd.)

[10] G. Salton, Automatic Phrase Matching, Information Storage and Retrieval, Report ISR-8, to the National Science Foundation, Section V, Harvard Computation Laboratory, December 1964.

[11] G. Shapiro, Statistical Phrase Processing, Information Storage and Retrieval, Report ISR-9, to the National Science Foundation, Section VII, Harvard Computation Laboratory, August 1965.

[12] R. T. Dattola and D. M. Murray, An Experiment in Automatic Thesaurus Construction, Student Report, June 1967.

[13] G. Shapiro, Processing of the Concept Hierarchy, Information Storage and Retrieval, Report ISR-7, to the National Science Foundation, Section V, Harvard Computation Laboratory, June 1964.

[14] M. Razar and G. Shipiro, Hierarchy Set-up and Hierarchy and Concept-Concept Expansion Procedures, Information Storage and Retrieval, Report ISR-9, to the National Science Foundation, Section XV, Harvard Computation Laboratory, August 1965.

[15] G. Blomgren, A. Goodman and L. Kelly, An Experimental Investigation of Automatic Hierarchy Generation, Information Storage and Retrieval, Report ISR-11, to the National Science Foundation, Section VIII, Department of Computer Science, Cornell University, June 1966.

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

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