VII. DICTIONARY CONSTRUCTION AND UPDATING

Claudine Harris

1. Introduction

The dictionaries necessary for the SMART system were described in Information Storage and Retrieval, Report No. ISR-7, Sec. III. They consist of word stem and suffix lists which provide for each stem a set of semantic concept numbers and syntactic values, a hierarchical structure of concepts, a list of syntactic criterion trees (phrases), a list of significant word pairs (phrase names), and a simulated vacuous dictionary or null dictionary. The thesaurus of word stems and concept numbers is available as an alphabetic list of all word stems with corresponding concept numbers and syntactic codes, and as a numerical dictionary of all concept numbers with the word stem to which they correspond.

Construction of the various dictionaries was carried out manually to the point where experimental lookups could be made to test the options of the SMART system. The following parts describe the procedures used in building the dictionaries and in revising them. Word stems were classified into broad groups first, and the output of the first lookup operations was used to refine the classifications. Constant consultation was maintained with the members of the group, leading to many valuable revisions.

2. Formation of the First Dictionaries

The first classifications were made on an entirely intuitive basis following the outline of concepts given in Information Storage and Retrieval, Report No. ISR-7, Sec. III. Out of the 405 documents, a representative group of 48 was selected and most of their technical vocabulary, as well as some of the nontechnical words, were assigned concept numbers and entered into the thesaurus. These selected documents were used in the early phases to test the various SMART routines; the automatic lookup yielded a list of "words not found" which could then be considered for addition to the dictionary.

The same procedure was followed for the remainder of the collection. The first lookup of the whole document collection was made after most words had been included in the thesaurus. The words not yet found in the thesaurus amounted, in general, to not more than a dozen per document, including abbreviations, proper names, etc. As new words were added to existing concept numbers and new concepts were created, a graph of the hierarchy of concepts was maintained up to date and constantly referred to. Invaluable contributions were made by the various members of the group in evaluating the hierarchy and by the individual word stems associated with each concept number in the numerical dictionary.

At this stage, there were still a number of "words not found," and reference was made to the alphabetic null dictionary to ascertain that the "words not found" did not occur with a frequency of four or more.

The versions of the dictionaries arrived at contained 1,944 word stems corresponding to 418 concepts; this original version was used in the first experimental lookups of the SMART system. Various updating procedures described below led to a second version which was used in the more extensive later experiments.

3. The Null Dictionary

The null dictionary is described in Information Storage and Retrieval, Report No. ISR-7, Secs. III and XI. All distinct word stems, after right-to-left suffixing, are assigned unique concept numbers. All words, which occur in a common word list, and all word stems with a frequency of occurrence of less than four are ignored.

In the 405 document abstracts used in these studies, the null thesaurus generated by the THES program contained 2,864 distinct word stems after suffixing and elimination of common words. With the addition of a frequency restriction to four occurrences or more, the dictionary size was brought down to a total of 898 distinct "concepts" (word stems).

The output null thesaurus thus generated was available in two forms for comparison with the SMART dictionaries: the 2,864 distinct noncommon word stems listed by frequency of occurrence, and the alphabetic list of noncommon word stems that appeared more than four times. These two outputs were used to update the SMART dictionaries. Excerpts from the null outputs are shown in Figs. 1 and 2.

Word Stem	First Suffix Detected	Frequency
solv	ing	22
spe	ed	22
success	ive	22
suit	able	22
variable	s	22
ampli	fying	23
arithmet	ic	23
communic	ation	23
mechan	ism	23
term	s	23
count	er	24
device	s	24
fast	er	24

Excerpt from the Null Thesaurus Frequency List

Figure 1

Word Stem	First Suffix Detected	"Concept" Number
slow		749
small		750
so-call	ed	751
solut	ion	752
solve		753
solv	ing	754
sort		755
source	s	756
space		757
special-purpose		758
specific	ations	759
speci	al	760
spectr	al	761

Excerpt from the Null Alphabetic List of Word Stems with Frequency Larger than Four

Figure 2

4. Updating the Dictionaries

A. The Null Alphabetic List

The alphabetic list of frequent word stems from the null dictionary was used to ascertain that no "important" words had been omitted from the dictionaries. The word stems from the alphabetical list that were omitted are, in general, words of nontechnical usage which were not thought likely to make a significant contribution to retrieval for technical subject matter.

Understandably, the choice of nontechnical words to be included in the dictionaries may be open to criticism, and we are aware that decisions could have been made at this stage on a less-than-adequate basis.

B. The Null Frequency List

A partial list of high-frequency word stems supplied by the null lookup is given in Table 1. Each stem to which a concept number is assigned by the null thesaurus is given without the suffixes that may give rise to it. All stems are listed in order of decreasing frequencies of occurrence down to 37 occurrences. A few words of lesser frequency are appended as a matter of interest.

Since the null thesaurus is constructed with right-hand suffix splitting, separate stems may be created from the variant grammatical forms of certain words. Typical cases are "process," "require," and "use":

Word	Frequency	Word	Frequency
comput	508	technique	83
system	263	automat	83
describ	233	anal	83
method	209	numb	77
digit	186	magnet	77
discuss	154	programm	73
oper	139	switch	72
circuit	130	control	71
program	127	develop	69
machine	124	applic	69
use	123	consider	67
gener	121	time	65
funct	112	form	65
design	106	differ	64
equ	102	solut	63
used	99	simul	62
log	98	perform	60
problem	94	inform	60
mem	94	ele	59
dat	88	electron	58
analog	86	stor	57
giv	85	present	56
point	56	line	44
langu	54	character	44
err	52	shown	42
bas	52	obtain	42

Most Frequent Noncommon Words

TABLE 1

Word	Frequency	Word	Frequency
network	51	ord	41
transl	50	bin	41
result	50	occur	41
output	50	matrix	40
set	49	volt	38
transist	48	code	38
stud	48	process	37
input	48	require	35
calcul	48	proces	35
state	47	pulse	34
mean	46	rel	33
speci	45	complex	33
core	45	routine	32
using	44	random	31
type	44	sign	29
requir	44	transistor	16
proced	44	processe	12

TABLE 1 (continued)

Stem	Suffixes	Frequency
process	ing, ed, or, etc...	37
proces	s	35
processe	s	12
requir	ing, ed	44
require	s, ment	35
use	s, ful, less	123
used		99
using		44

It is seen that the separate stems give rise to misleading frequency counts. For example, the total frequency of the common forms of "use," not including "usable," "user," and "usage" with frequencies of 1, 3, and 3 respectively, comes to 266, thus making it the second most common word stem in the documents.

C. "Use" and "Discuss"

A word, as common in the English language as "use," can be expected to appear in the requests submitted to the SMART system, as indeed it does. In a total of 24 requests prepared by a variety of people, including both requests in the natural language and in outline form -- that is, without complete grammatical sentences, the following distribution of "use" and "discuss" was found.

	Number of Requests	Occurrences of "use"	Occurrences of "Discuss"
natural language	14	8	6
outline form	10	0	0

This is sufficient to suggest that such everyday nontechnical terms be labeled as "common" words. With this in mind, the request and document correlations obtained with the first dictionary lookup were searched for ambiguities introduced by everyday words. When documents that were not relevant to a request were assigned a spurious high correlation, it was found that the largest contribution to the correlation came from everyday noncommon words such as "use," "discuss," "procedure," "method," etc...

Hence, one of the first modifications undertaken in the construction of the second dictionary was to eliminate all such words by classifying them as "common." In so doing, the concept number for "discuss" which corresponded to 37 ways of saying "this is what we are talking about" was included in the "common" list. Words from the high frequency table affected by this change were "discuss, give, develop, consider, present, study."

The words "method, technique, procedure" were also made "common," as were their 16 synonyms. In this collection, the words "handling" and "processing" are technically significant, as in "data processing," and their stems were retained.

5. Refining the Numerical Dictionary by Concept Concordance

The erroneous correlations introduced by "common" words during the first lookup gave an indication that similar spurious correlations could arise from technical terms if the numerical concepts which included them had been defined ambiguously. For example, the word "generate" was originally assigned to the concept called 80MAKE together with 16 other words meaning "construction." The request RANDOM NUMBS, which reads: "How can one generate random numbers efficiently? Which pseudo-random sequences offer long periods and ease of generation?" obviously does not need to be correlated with the many occurrences of the words "make" and "made" (total frequency 62).

The analysis of ambiguous concept numbers was made by studying a concept-document concordance and referring back to individual documents to ascertain the specific usage of the concepts in each case. Results were tabulated and the various words reassigned, if necessary. New concepts were often created in this process, and the hierarchical structure revised. Typical concept analyses are given below.

A concept number was selected for analysis either because it was visibly ambiguous (80MAKE), or because it was very frequent (168HIE), or because it appeared in a request with poor document correlations (101TAG), or for a combination of reasons. The concordance between occurrences of a concept number and the documents was at first established by visual scanning of the document vector output; a program now exists, however, to produce the concept concordance automatically. Future updating of the dictionary toward a third version,

as well as the formation of a new thesaurus for a different document collection, will be much aided by this analysis program.

A. Analysis of 101TAG

The concept 101TAG which contained 13 words meaning "designate, identify" was chosen for analysis because it occurred in two requests and seemed ambiguous. The word list attached to this concept included: call, designate, identi, index, indicate, label, mark, name, point, signal, sign, subscript, tag. The concept occurred in 94 documents, with the following distribution of significant terms:

Term	Frequency	Number of Documents
index	17	7
signal (pulse)	20	14
identify	6	4

All other terms under this concept occurred a total of 91 times, accounted for almost exclusively by "pointed out, indicated, called..."

As a result, the words "indicate, call, name, designate" were made into common words; the words "sign, signal, mark, point" were removed from 101TAG but kept in their other respective concepts; "identi" was moved to the concept number for "recognition." Only "index, label, subscript, tag" remained under 101TAG.

B. "Point" and "Pointed out"

The previous analysis led to a study of the occurrences of "point."

Meaning of "Point"	Frequency
mathematical usage	12
location	24
"pointed out"	15
purpose	0

Only the meanings of location and mathematics need be retained for "point"; the meaning of "purpose" can be ignored; the words "pointed" and "pointing" can be made common.

C. Analysis of 168HIE

The concept 168HIE was an ambiguously wide concept meaning "ordering, or classification." It included: classi, grade, hierarchy, interval, level, order, sequ, serial, step, succession, successive, tree. The concept occurred in 97 documents and in two requests with the following distribution of terms:

Term	Frequency
order	41
step, interval, level	31
sequence	23
sequential	15
classify, grade	14
successive	13
serial	7
tree	1

The word "order" was further studied as to distribution of meanings:

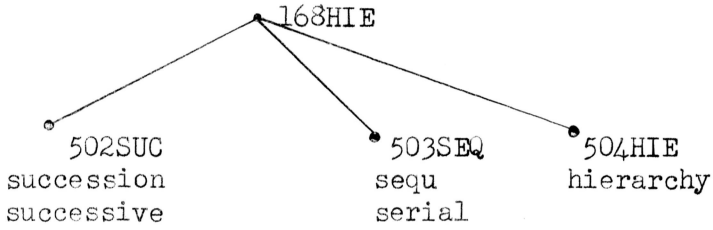
Meaning of "Order"	Frequency
ordering (sequence)	11
mathematical usage	10
command	6
purchase	2
"in order to"	12

Because of the frequent significant uses of "order," it is not possible to eliminate this word by making it into a common one. However, we can ignore the meaning of "purchasing" and assign "order" to only three major concept numbers.

It was thought best to separate "successive" from "serial" to allow for the phrase "successive approximation." The word "tree" was moved to an existing concept for words descriptive of graphs.

The final assignment of words from 168HIE is shown in the following excerpt from the concept hierarchy, with the more specific words placed under the general concept of "classification":

classi
grade
interval
level
order
step
strati



Other types of ordering, such as "reversal" and the phrase "random number," were also introduced under 168HIE.

D. Analysis of 80MAKE

The ambiguous concept 80MAKE contained the words: arrange, build, built, construct, create, establish, fabricate, fashion, form, generate, install, made, make, produce, product, shape, work. The concept occurred in two requests and in 165 documents with the following distribution of terms which shows how the words were regrouped.

Term	Frequency
create, generate, produce	
production	74
make, made	62
form (geometry)	53
build, built, construct, fabricate	41
work (task, action)	21
actuate, establish, install	8
shape (geometry)	8
product (goods, result, math)	7

The words "make, made" became common; all others were separated as indicated above.

6. The Phrase Dictionaries

Construction of the phrase dictionaries is not yet as complete as that of the word stem and concept dictionaries. To date, 56 distinct phrase names are recognized. They are assigned variously to existing concept numbers or to concept numbers of their own and tied into the numerical concept hierarchy. It is not possible to list the exact number of English phrases which can be recognized, since a phrase name can correspond to many combinations of word stems. For example, the phrase name LNGTRA, concept number 303 and means "language translation," can be formed by three pairs of concepts: 98TRSL and 102LNG, 98TRSL and 35LANG, and 98TRSL and 114TEX. The multiplicity of possible corresponding phrases in English can be seen from a list of the word stems under these concept numbers:

98	102	35	114	
transcribe	interlingu	Cyrillic	abstract	literature
transcript	language	English	article	page
translate	lingu	French	bibliograph	paper
word-for-word		Morse	catalog	passage
		Roman	copy	report
		Russian	document	text

The phrase dictionary contains 99 such combinations corresponding to 56 distinct phrases. Since the phrases are concept pairs, rather than word pairs, many worthless pairs of words co-occurring within a sentence will be identified as phrases. For example, the phrase name SUCAPP, which is intended to mean "successive approximations," is made up of co-occurrences of 502SUC (succession, successive) and 106NQU (approximate, estimate, incomplete). While the intended noun phrase is included in this specification, sentences such as "We estimate that a succession of experiments is necessary" or "Successive attempts to refine the translations still yield an incomplete text" are also included. In the first sentence, SUCAPP is obtained by the pair "estimate succession" and in the second sentence by the pair "successive incomplete."

Such ambiguities are resolved by syntactic analysis as described in Information Storage and Retrieval, Report No. ISR-7, Sec. VI. Briefly, the SMART system provides for the syntactic analysis of selected sentences by means of the Kuno Syntactic Analyzer (Mathematical Linguistics and Automatic Translation, Reports NSF-8 and NSF-9) and for the matching of these sentences with a dictionary of preanalyzed syntactic phrases called "criterion trees." Criterion trees are syntactic phrases with specified

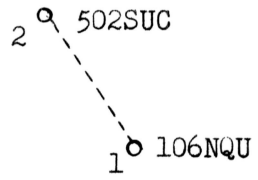
concept numbers and syntactic dependency relations between the included concepts.

Returning to the example above, the dependency relations which can be assigned, are, in general, noun phrase, subject verb, verb object, and subject object. In the criterion tree dictionary SUCAPP has been assigned only a "noun phrase" criterion tree, which will not match the verb object structure "We estimate that..." nor the subject object relationship of "Successive attempts...yield an incomplete text." In this manner, the spurious phrases "estimate succession" and "successive incomplete" will not be rendered as "successive approximation."

The syntactical "criterion tree" dictionary now contains 108 trees corresponding to the 56 phrase names and innumerable English sentences. The choice of tree structures for a given phrase name was made by considering the words within the concept pairs and the possible sentences in which they might be used. At this time, there are 14 distinct tree types, but most phrases are given only the most common syntactic subgraphs for adjective noun or verb object structure. As an example, the following example gives the allowed structures for the phrases LNGTRA and SUCAPP.

Phrase Name	Tree Type Allowed	Example
LNGTRA	adjective noun	language translation
LNGTRA	verb object	we translate the documents
LNGTRA	noun complement clause	the documents are suitable for translation
SUCAPP	adjective noun	successive approximation

An additional feature of the criterion trees is that the syntactic dependencies are directed so that a given structure is not reversible. The simple adjective noun tree assigned to SUCAPP consists of two nodes indirectly connected as shown below.



Criterion Tree
SUCAPP

Node 1 must have the concept number 106NQU, and node 2 must have the concept number 502SUC. The numbering of the nodes indicates the dependence of node 2 on node 1; that is, the adjective relationship between node 2 and node 1 is fixed. This specification accepts "successive approximation," but not "approximate succession"; "we translate the documents," but not "we have documented the translation."

Evaluation of the dictionary specifications is continuing in the light of continuing computer experiments.