NASA Contractor Report 4512

Machine Aided Indexing from Natural Language Text

Status Report

June P. Silvester, Michael T. Genuardi, and Paul H. Klingbiel

CONTRACT NASw-4584
MARCH 1993

NASA Contractor Report 4512

Machine Aided Indexing from Natural Language Text

Status Report

June P. Silvester, Michael T. Genuardi, and Paul H. Klingbiel
RMS Associates
Linthicum Heights, Maryland

Prepared for
NASA Scientific and Technical Information Program
under Contract NASw-4584

National Aeronautics and
Space Administration

Scientific and Technical
Information Program

1993

This publication was prepared by the NASA Center for AeroSpace Information,
800 Elkridge Landing Road, Linthicum Heights, MD 21090-2934, (301) 621-0390.

TABLE OF CONTENTS

FIGURES

PREFACE

This report describes the machine aided indexing (MAI) system that was developed for the National Aeronautics and Space Administration (NASA) Scientific and Technical Information Program at the NASA Center for AeroSpace Information (CASI). The NASA Lexical Dictionary (NLD) MAI system automatically produces candidate, NASA- controlled-vocabulary, index terms from any designated natural language text input. Development of the NLD MAI system was begun under contract NASw-3330 by the Planning Research Corporation/Government Information Systems and was concluded by RMS Associates under contracts NASw-4070 and NASw-4584. The period of performance covered by this report is from 2 November 1981 to 31 October 1992.

We would like to acknowledge the help given to our project by the people at the Defense Technical Information Center (DTIC) who shared their expertise, their programs, their files, and their controlled vocabulary with us when we began NASA's NLD MAI project. Without their cooperation the construction of the NLD would have been more difficult and more costly. We also acknowledge the encouragement and opportunities that Herman Miles of PRC and John Wilson of JTT at NASA Headquarters have provided. The MAI project was undertaken because of their foresight.

EXECUTIVE SUMMARY

The NASA Lexical Dictionary (NLD) machine aided indexing (MAI) system was designed to (1) reuse the indexing of the Defense Technical Information Center, (2) reuse the indexing of the Department of Energy, and (3) reduce the time required for original indexing. This was done by automatically generating appropriate NASA Thesaurus terms from either the other agency's index terms, or, for original indexing, from document titles and abstracts. The NASA STI Program staff devised two different ways to generate Thesaurus terms from text. The first group of programs identified noun phrases by a parsing method that allowed for conjunctions and certain prepositions, on the assumption that indexable concepts are found in such phrases. Results were not always satisfactory and it was noted that indexable concepts often occurred outside of noun phrases. The first method also proved to be too slow r the ultimate goal of interactive (online) MAI. The second group of programs used the Knowledge Base (KB), word proximity, and frequency of word and phrase occurrence to identify indexable concepts. Both methods are described and illustrated. Online MAI has been achieved, as well as several spinoff benefits, which are also described.

INTRODUCTION

Purpose

The NASA MAI project has had two main purposes. Tne first was to minimize the re-indexing of documents already indexed by another agency. When the project was begun, approximately half of the report literature that was added to the NASA STI database each year had been previously cataloged, abstracted, and indexed by another agency. These reports, received in machine-readable form, are converted automatically not only to NASA's data elements, but also to posting terms selected from NASA's controlled vocabulary through either Subject Switching or machine aided indexing.

The second purpose was to reduce the time required for original indexing by providing the correct form of a term, the correct technical term, or a synonymous NASA term for a concept expressed by an author, and by providing a check list of potentially indexable concepts. This is done by generating a set of candidate, authorized, NASA terms for indexers to accept or reject. Such terms are generated by processing natural language text, primarily titles and abstracts, through MAI.

Significance

The NASA MAI system is significant for its versatility. New applications continue to keep the system experimental. At the same time it serves as an efficient and accepted aid for indexers, for proofreaders, for the Thesaurus Lexicographer, and for searchers. It has unrealized potential for further assisting indexers and searchers, and for helping other agencies that want an operating MAI system of their own. The NASA system is also significant for pioneering not only an operating MAI capability but for making it available online in an interactive mode.

Definition

We quote from NASA CR-3838: "A lexical dictionary has been defined in several ways. Paul H. Klingbiel, who initiated the NLD defines it two ways...as 'a phrase structure rewrite system' and as 'a matrix.'" Roxanne Newton - second director of the project - defined the NLD (in the "System Overview" written for internal use) as "a translation device." The different descriptions represent different points of view. To a mathematician, a lexical dictionary is a matrix; to a linguist, it's a grammar; to an accountant, the system resembles a spreadsheet; to those dealing with operating systems, the lexical dictionary is a translation device; but to the indexer, the lexical dictionary is a tool.

This report addresses itself primarily to the indexer's definition --that is, that the lexical dictionary is an indexing tool. Secondarily it demonstrates that the lexical dictionary is a matrix, and finally it discusses the linguistic grammar aspect of the NLD system.

Scope of the NLD

The scope of the NLD project keeps widening as more applications are found for its programs, files, and by-products. When the project was begun, the only use of the NLD was to switch DTIC's indexing of technical reports automatically to NASA's controlled vocabulary. When that phase was complete, file and program construction were repeated for reports indexed by DOE. At the same time, a file was being built to handle natural language input from titles, abstracts, or any other designated source. We continue to find new uses for the NLD as its capabilities expand.

For example, the program that accesses the NLD puts out a report that lists "Words not found." This list identifies words and phrases that may need to be added to the NLD Knowledge Base (KB). It also identifies words that cannot be translated by the KB because they are misspelled. These words are now listed for proofreaders in an online display, at the end of the line in which the "not found" word occurs. A quick look at the list to the right of a paragraph is an efficient way to spot misspellings. Also, because the Input Processing System is an online system, correcting misspellings is much like correcting an error on a word processor. The NLD spelling check feature is better than a commercial system because KB entries are generated from NASA literature, and therefore contain many technical words that a commercial system would not include.

Other NLD programs have been used to identify needed new Thesaurus terms and Use references. For a time, the KB also provided NASA posting terms for Library of Congress records that were entered into the NASA database. As always, we stress that the NASA MAI system is not machine indexing, but rather a tool for indexers.* Candidate, authorized NASA terms are suggested and indexers accept, reject, and/or add to the MAI output.

The DTIC to NASA Subject Switching portion was described in detail in NASA CR-3838. DOE to NASA Subject Switching is virtually identical. This report documents NASA's natural language machine aided indexing, both the original system and its second generation system available in an interactive mode online. It also identifies some of the peripheral uses for MAI programs and the KB.

*The consideration of MAI as machine indexing (MI) in certain environments always leads to stimulating discussions.

Preliminary Results

MAI was developed at the NASA Center for AeroSpace Information in a production environment. Measurements of results were devised to be non-disruptive of the regular work flow. The following observations have been made:

- The indexing staff has decreased from 8 to 5 people.
- The workload has not decreased.
- Indexing is more consistent between indexers than it was before MAI. (This was noted by the person who has trained 80% of the present staff.)
- Fewer errors of omission are made. (Also noted by the trainer.)
- Less research time is required because of the expert advice provided by MAI as to appropriate technical terms.

It is reasonable to conclude from the above that the indexers, supported by the MAI system, have been able to maintain and even improve indexing quality, and at the same time increase production.

Presentation

This report provides an overview of the NASA MAI system. It also includes a step-by-step account of the development of both the original and the second generation system, a description of system maintenance, the resources required for MAI implementation at CASI, and finally the benefits, problems, and future of the NASA Lexical Dictionary system.

Project Personnel

The MAI project was begun and has been guided by Paul H. Klingbiel, Consultant. It is currently under the supervision of Project Coordinator June P. Silvester and Analyst Michael T. Genuardi. Also participating in the work were the entire indexing staff and numerous Programmer Analysts, including especially Carl N. Collins and Robert W. Egge.

HISTORY

DTIC's Role

Paul Klingbiel, first director of the NLD Project, had been active for 18 years in linguistic research at DTIC. While there, he had initiated a Lexical Dictionary (LD) which became part of DTIC's machine aided indexing system. The present DTIC LD had its origins in the Natural Language Data Base (NLDB) which was established between 1974 and 1979 at DTIC, then the Defense Documentation Center (DDC). The core vocabulary of the NLDB was the DDC thesaurus with the omission of related and hierarchical terms. Natural language phrases (with a maximum length of four words) were added from MAI production runs when they did not match an entry in the NLDB. The available manpower was not sufficient to cope with the large number of phrases produced by MAI. Nevertheless, in approximately 4 years about 250,000 natural language phrases were added to the core terms already in the NLDB. A preliminary edition of the DDC Retrieval and Indexing Terminology published in 1975 listed 184 natural language phrases containing the word "ship" or "ships." Projections indicated that the NLDB would at least double in size before the number of new candidate phrases substantially decreased. A final total of a million phrases was quite possible. Consequently, building an NLDB was abandoned in favor of a new, more compact structure call the Lexical Dictionary (Klingbiel, 1985).

After retiring from DTIC, Klingbiel agreed to organize machine aided indexing at NASA. Copies of DTIC's programs and prints of their LD were obtained and studied, but could not be used directly because computer languages and equipment at the two agencies are not compatible. DTIC's programs were written for a UNIVAC mainframe and sent to NASA in COBOL, while NASA's programs were written in PL1 for an IBM mainframe. A tape of DTIC's LD was also obtained. When this file was inverted, it provided information on how the NASA Lexical Dictionary (NLD) Knowledge Base (KB), (which was founded on the DDC LD structure*), could translate DTIC posting terms. It was helpful, as well, in identifying natural language phrases that could be translated into NASA posting terms. The DDC LD was built from MAI production output. The NLD KB has been constructed from a variety of sources, but is now largely being expanded from analyses of text targeted to specific thesaurus terms, as explained in the section on the KBB Text Analysis Tool (p. 44). Whatever procedure is used, the intent is to build a KB sufficiently comprehensive to translate Access-2 output in such a manner that indexers input is largely editorial.

* Its modification and current status are described in the Knowledge Base Building Procedures section.

NASA KWOC and Data Entry

Klingbiel began the KB with a list of NASA posting terms in a special Keys Words Out of Context (KWOC) format. A KWOC listing had been used at DTIC to review and correct inconsistencies that had entered into their Natural Language Database. By starting the KB with a KWOC printout of all of NASA's posting terms and Use references, the problems experienced at DTIC were avoided. It was determined later that an alphabetized list of NASA posting terms would have worked just as well. The use of the KWOC was described in detail by Silvester, Newton, and Klingbiel (1984) in NASA CR-3838. Suffice it to say here that each authorized posting term and Use reference that appeared in the NASA Thesaurus was coded and entered with an appropriate logic code into the KB.

Completion of this phase had two results: (1) the capability for automatically translating, i.e., Subject Switching (SS), any DTIC posting term that exactly matched an authorized NASA term, and (2) a decision to separate Subject Switching (SS) files and procedures from those files and programs that translate natural language words and phrases to authorized NASA terms. The SS of DTIC to NASA terms became operational in June 1983 and was fully described in NASA CR-3838. During the following year a similar SS project was untertaken for translating the authorized posting terms of the Department of Energy (DOE). This was a much larger task and while never totally completed, it is able to translate virtually all of the DOE terms that NASA encounters. The omissions are largely highly specific atomic energy terms and table entries for linked DOE terms.

This report describes the machine aided indexing (MAI) system used at the NASA Center for AeroSpace Information. Input is natural language text. Output is a set of NASA authorized posting terms from its controlled vocabulary as listed in the NASA Thesaurus, Thesaurus Supplements, and Update Sheets.

The NASA electronic Input Processing System (IPS) divides the file series that are cataloged, abstracted, and indexed into two sections: the alternate files (IPS-ALT) which are entered either infrequently or irregularly, and the primary IPS records, which are entered daily and include STAR.

MAI of natural language text was begun using the NASA Recognition Dictionary (NRD), the Machine Phrase Selection (MAPS) program, and the original Access Routine, Access-1, to identify indexable concepts. This system became operational in an IPS-ALT file in August 1986. MAI with the NRD, MAPS, and Access-1 become available as an online, interactive system for primary IPS documents without abstracts in October 1988. Documents with abstracts took too long for online use of the system. They required an average of a minute and a half, which was an unaccepably long wait for MAI-suggested terms. Access-2, designed to eliminate the need for parsing and to shorten processing time, became operational in overnight batch mode in May 1989 for STAR analytics, in daytime batch mode for other STAR documents in March 1990, and for all of the "primary document series" in an online, interactive mode in June 1990. The wait time had been reduced from about 90 seconds to 6 or 7 seconds.

SYSTEM DESCRIPTION

Significance Within the Larger Scope

As stated earlier, the NLD is a tool; however, looked at as part of a larger system, the NLD is more than that. Its use has altered indexing procedures; improved quality control; promoted creativity; stimulated communication, not only between internal groups, but also between organizations; enhanced NASA Thesaurus construction; and brought about a new awareness of shortcomings and strengths of various thesauri and the need to improve terminology standards within the government. The NLD has become a tool, not only for indexers, but also for proofreaders, lexicographers, standards groups, and those who search the database.

System Functions

The NASA MAI system generates NASA Thesaurus (controlled vocabulary) terms from any designated natural language input. This is usually, but not always, technical report titles and abstracts. Input also may include another organization's controlled vocabulary terms. (This is totally different from Subject Switching which is fully described in NASA CR-3838.) Indexers review the output NASA Thesaurus terms and either accept or reject them. Indexers may also add Thesaurus terms not generated by MAI. Feedback is encouraged to improve the file records that provide the input-to-output translations.

MAI with Access-1

System Components. The components of NASA's first MAI system (which is a second generation of the MAI system initiated by Klingbiel at DTIC) are:

- o the NASA Recognition Dictionary (NRD);
- o the Machine Phrase Selection (MAPS) program;
- o the data file - originally referred to as the Phrase Matching file and now called the Knowledge Base (KB);
- o application programs that indicate the input words or strings to be processed and the kind of processing wanted, i.e., natural language MAI or Subject Switching; and
- o Access-1, a program that manipulates the language input, matches, the output against the KB, and returns the candidate NASA Thesaurus posting terms to the application program for output to the user.

These components are described in greater detail below.

The NASA Recognition Dictionary (NRD). This is a large file of data that was originally created at DTIC and modified by NASA for use at the Center for AeroSpace Information (CASI). At this writing, it consists of 139,269 unique, English language, all-alpha words that constitute the the values for the key fields of these records. Three other fields exist for each NRD record: (1) the word category; (2) the category assignment origin flag; and (3) the count.

o The category provides syntax that is used in machine phrase selection.

o The flag identifies the origin of the category assignment. If the category assigned is in need of being changed, it is important to know what the flag is and what it means.
  * Words flagged with an asterisk were categorized by DTIC.
  # Words flagged with a pound sign are authorized NASA terms or parts of authorized terms.
  ? Words flagged with a question mark were assigned Category 06 automatically, with limited or no research, when a large number of words, categorized as stopwords (01) by DTIC, were added to the NRD as context-sensitive, weak nouns.
  b Words flagged with a blank were words whose categories had been changed by STI Program staff in response to perceived problems.

o The count field provides information on word occurrence in text that has been processed through MAI and the NRD by NASA. Each time a word is looked up in the NRD, the program raises the count for that word by one. For more information on the NRD, see Appendix A.

Machine Phrase Selection (MAPS) Package. Five programs were written for the MAPS package:

o An applications program that provided the following:
  - identification of the input text for MAI;
  - a list of words in the text that were not found in the NRD;
  - a file of "good" phrases, i.e., phrases that ended with word categories D (19), G (21), H (16), J (22), N (03), or Z (06);
  - a file of "bad" phrases, i.e., phrases that contained none of those categories; and
  - a list of phrase formats, i.e. the letters that marked each word and comma in a phrase (such as AN, AZN, JG, ZHZ) and served as a "trace" for the analysts.

  See components below for additional information on the applications programs and see Appendix B for grammar rules used to delineate phrases for MAPS input.

o The MAPS program, stored as a procedure. This consisted of several subroutines and functions whose main purpose was to analyze the input text and break it down into phrases.

o A print program written specifically to list the phrases that did not fall into the "good" phrases group.

- o A program that produced three reports from the file of "good" phrases:
  - sorted by phrase
  - sorted by trace format
  - a summary of trace format variations by frequency

- o Access-1, also stored as a procedure, analyzed the "good" phrases for multi-word strings and single words that might express an indexable concept. These were matched against the KB. Outputs from Access-1 were candidate terms, completely matched phrases, partially matched phrases, and phrases that did not match any key at all in the KB. See below for additional information on Access-1.

The Data File. The data file for MAI, the KB, was originally a matrix of four fields: The initial word of the input, the final word of the input, the posting term(s), and the logic code. In the summer of 1982, the NASA systems were changed over from an IBM-360/65 computer using a Multiprogramming Variable Tasks (MVT) operating system to an IBM-4341 computer using a Multiple Virtual Storage (MVS) operating system.* The Virtual Storage Access Method (VSAM) required that each record have a a unique field, or key that serves as an address for the record. The construction of this key reduced the number of fields from four to three: (1) the key; (2) the posting term(s); and (3) the logic code.

Key: This unique field was constructed (Silvester, Newton, and Klingbiel, 1984) by concatenating the first and final input word fields. When an entry consists of a single word, the value of the final word field is null. At DTIC, this null word was marked with two zeros, because DTIC's computers sort zeros last. NASA's computers sort nines last. In copying the DTIC system, NASA's original MAI system also marked this null word with zeros. (This was changed to three nines when the system began to use Access-2.) All NASA Thesaurus posting terms and Use references are included as keys, as well as many additional natural language words and phrases that are the conceptual equivalents of the NASA Thesaurus posting terms.

When the key contains more than two words - i.e., three or more words or two or more words and ";00" - intermediate records, called continuation entries, are created to build to the multi-word key. The key of the first entry consists of the first two words. Each continuation entry adds one more word to the key until the key contains the entire phrase. In practice, it was found that these intermediate records were sometimes forgotten, and therefore a program was written to check for, and if necessary create, the continuation entries.

* Later a second IBM-4341 mainframe was added to the CASI system, and by 1987 both of these had been replaced by IBM-4381 mainframes using the MVS/XA (Extended Architecture) operating system.

Posting Term(s): When the key contains the complete phrase to be translated, the posting term field may contain:

(1) One or more NASA Thesaurus terms that express the same concept, Terms found are returned to the indexer for acceptance or rejection.

(2) Two zeros. Two zeros in the posting term field mean either that no appropriate NASA translation is available, or that the concept expressed by this key is usually not indexed to by human indexers when the word or words in the key are encountered in natural language text.

(3) Continuation symbol. The continuation symbol tells the computer to look for an additional or a null (00) word to find a key that will translate. Originally there was a progression of symbols (*, **, %, %%, %%%) used in the posting term field to indicate continuation. These are explained in NASA CR-3838 (Silvester, Newton, and Klingbiel, 1984) The current system uses only a single asterisk (*) as illustrated in the sample below. If a translation is wanted for the key of a continuation entry, a semicolon and two zeros are added at the end of the key, to creat a unique key, and the appropriate translation then is entered in the posting term field.

Logic Code: The logic code field in the original MAI matrix was a one character code that indicated how the key was to be processed by the Access-1 program. Keys that had only a single word (followed by 00 for the null final word of the input) were assigned one of the following logic codes:

E - (Equal) when the key translated to a single posting term that was the same as the key.

C - (Change) when the key translated to a single posting term that was different from the key.

L - (List) when the key translated to multiple posting terms that should be used together.

0 - (Zero) when the key had no translation or none was wanted.

When the key had more than one word, or when it had a word or words that also appeared in the initial position of another key, the logic code was a T for Table. A Table occurs when the first word of the input is context sensitive and requires another word or words to clarify the concept. The T told the computer that, when it encountered the first word, a second word must be sought. If the text did not contain any of the additional words that appeared in the matrix (Knowledge Base), then the system defaulted to the first word followed by the null word value of ";00" and, if that entry could be found, returned the posting term(s) found for that entry.

Some examples of logic codes, keys, and postings are shown below:

Logic Code	Key	Posting Term(s)
E	Aborigines;00	Aborigines
C	A-star;00	A-stars
L	AC-209;00	Actinium,Isotopes
0	Academy;00	00
T	Airport;traffic	*
T	Airport;traffic;control	*
T	Airport;traffic;control; personnel	Air traffic controllers (personnel)
T	Airport;traffic;control;00	Air traffic control
T	Airport;traffic;00	Air traffic
T	Airport;00	Airports

Three columns were reserved for recording logic codes, which were adopted from Klingbiel's original "lexical dictionary" format (Klingbiel, 1985). A logic code was assigned to each entry in the Knowledge Base. Codes recorded in the first column indicated that the Access-1 program was to use the Phrase Matching file. This was referred to as the operation of Access-1 in Mode 1. Codes recorded in the second column told the computer to access the DTIC-to-NASA Subject Switching file, while codes in the third column told the computer to access the DOE-to-NASA Subject Switching file. Subject Switching used Access-1 in Mode 2.

Applications Programs. The Applications program, different for each use of MAI, is described under MAPS, above. It handles input, some of the output, and interfaces with the Access program. It passes two things to Access-1: a code that indicates which file is to be accessed, and a a character string that is either a word or a phrase from which Access-1 constructs potential NLD keys for lookup in the KB.

Access-1. The Access routine was written as a modular program that is called by an applications program. It constructs search keys from the input textual strings, looks them up in the Knowledge Base, and returns thesaurus terms to the applications program.

Text Processing with Access-1. Processing of input text by Access-1 is done in two modes. Mode 2 is used for Subject Switching and was fully described in NASA Contractor Report 3838 (Silvester, Newton, and Klingbiel, 1984). Mode 1 is used for Machine Aided Indexing of natural language using the NASA Recognition Dictionary and the MAPs method of selecting input words and phrases. Mode 1 processing of input text by Access-1 is done as follows:

o The logic code associated with the first word in the input phrase was located in the Phrase Matching file (KB). This logic code controlled the way the word was processed.

If the word had a T logic code, a potential KB key was formed by adding a semicolon and the next word in the input character string (phrase). This key was then looked up in the KB and a match was attempted.

- o If the search key matched a key in the file that translated to a posting term, that posting term was returned.
- o If the search key matched a file key that contained a continuation character (*, **, %, %%, or %%%) in the posting term field, then the next word from the input phrase was added to the end of the search key. A match with a key in the file was again attempted.
- o If no match was found for the search key, then the final word of the search key was replaced with the next word in the input phrase and another attempt was made to match the search key with a key in the file.
- o If all subsequent words of the input phrase were tried without finding any match, then a ;00 was added to the combination of words already found posted to a continuation character, and this search key was looked for in the file. (Access-1 requires the use of 00 instead of 999.)
- o If this search key could not be found, the final word (if any) before the null word was dropped and the resulting key was looked for in the KB. This process was continued until either a match was found or the first word ;00 was used as a search key.

The "T" logic processing was repeated for each word in the input phrase that had a T logic code.

If the word had a logic code other than a T, and if it had not been used with a T word, then a potential (KB) key was formed by adding a semicolon and two zeros (;00) to the word, and an attempt was made to find a match for this key in the file. Any key found was immediately translated to the NASA authorized posting term or terms that were in the matrix for that key.

Access-1 returned all of the matches made from the input character string (phrase) to the application program.

The following example illustrates a simple case of Access-1 processing for MAI. KB file entries needed to process the sample input, and related KB entries, have been extracted and are listed below.

Logic Code	Key	Posting Term
E	ENDURANCE;00	ENDURANCE
T	ENGINE;CONTROL	ENGINE CONTROL
T	ENGINE;DESIGN	ENGINE DESIGN
T	ENGINE;TESTING	*
T	ENGINE;TESTING:LABORATORIES	ENGINE TESTING LABORATORIES
T	ENGINE;TESTS	ENGINE TESTS
E	LABORATORIES;00	LABORATORIES
T	RESEARCH;AND:DEVELOPMENT	RESEARCH AND DEVELOPMENT
T	RESEARCH;FACILITIES	RESEARCH FACILITIES
T	RESEARCH;00	RESEARCH
T	TESTING;MACHINES	TESTING MACHINES
T	TESTING;TIME	TESTING TIME
T	TESTING;00	TESTS

Given the input phrase:
"Engine endurance testing research laboratories"
create an input array of the words and their associated logic codes as shown below. The logic codes are determined by a lookup process in the KB.

Input Array or Phrase	Logic Code
Engine	T
Endurance	E
Testing	T
Research	T
Laboratories	E

In the following Access-1 processing, references are made to the input array and the KB entries shown above.

Processing Description:	Outcome:
Logic code of first word is T. Create search key from first two words. Look up search key "ENGINE;ENDURANCE" in file.	Key not found.
Replace final word of search key with next word in input array. Look up search key "ENGINE;TESTING" in file.	Key found. Continuation symbol returned.
Add next word in the input array to the search key. Look up "ENGINE;TESTING;RESEARCH" in file.	Key not found.
Replace final word in the search key with the next word in the array. Look up search key "ENGINE;TESTING;LABORATORIES" in file.	Key found. Posting term "ENGINE TESTING LABORATORIES" returned.

No more words remain in array to be tried as the final word in a search key beginning with "ENGINE." End processing for "Engine."	
Move on to next word in array. The logic code for "Endurance" is E. Create a search key by adding ;00 and look up "ENDURANCE;00" in the file.	Key found. Posting term "ENDURANCE" returned.
Move on to the next word in array. Logic code is T. Create a search key by adding the next word in the array. Look up search key "TESTING;RESEARCH" in the file.	Key not found.
Replace the final word in the search key with the next word in the array. Look up search key "TESTING;LABORATORIES" in the file.	Key not found.
Replace the final word in the search key with ;00. Look up search key "TESTING;00" in file.	Key found. Posting term "TESTS" returned.
Move on to next word in array. Logic code for "Research" is T. Create a search key by adding next word in array. Look up search key "RESEARCH;LABORATORIES" in file.	Key not found.
No words remain in array to be tried with "Research." Replace the final word in search key with ;00. Look up search key "RESEARCH;00" in file.	Key found. Posting term "RESEARCH" returned.
Move to next word in array. Logic code is E, i.e., not a T. "Laboratories" has already been used in a previous successful match. End processing for "Laboratories."	
No more words in array. End processing.	

The final outcome of processing with Access-1 was that the input phrase "Engine endurance testing research laboratories" would be translated to the suggested NASA posting terms:

ENGINE TESTING LABORATORIES,
ENDURANCE,
TESTS, and
RESEARCH.

In addition to producing authorized posting terms for the indexer to review, the application program provided the MAI team with several reports that aided in building the KB. Each machine selected phrase, any corresponding thesaurus term or terms, and each phrase's grammatical or "trace" format expressed in alphas, were written out for the Analysts to study. The trace formats, such as "AN", "NZN", "NZPAN", and so on, provided a way of determining the effectiveness of the grammar rules and the NRD word categories. Other reports were used to measure the "match rate" of the MAI system and to spot KB deficiencies. The Match Rate and other measures are discussed in the section on "Results and Conclusions." The various reports are discussed and illustrated in the section on Knowledge Base Building, KBB Procedures Used With Access-1. The meanings of the trace formats are summarized in Appendix A, Table A-1.

Problems. One problem with our first MAI system was that it suggested many terms that human indexers would avoid. For example, they would avoid both "Endurance" and "Tests" because these terms lack specificity. Such terms are called array terms. NASA indexers use array terms rarely - only when no other term is broad enough for the content of the document being indexed.

Another problem that was noted was the frequency with which MAI suggested single word terms. Although approximately 40% of NASA's authorized posting terms are only one word long, (probably a holdover from the days when NASA indexed with uniterms), only 20% of the terms in an index set for a document are single terms.

The first of these two problems was addressed by changing most of the array terms' posting term fields to 00. In a few instances where it was felt that the array term would be used, an "at" sign (@) was inserted after the term to flag it for the indexer who reviews the output. For example, military aircraft@ is an array term that might be used by NASA's human indexers, for lack of a more specific term. Access-2 also reduced the number of broad terms that were suggested by MAI.

The second problem was reduced with the aid of a single word term assessment tool, which is described in the section on Procedures for Building the Knowledge Base and in Appendix D.

MAI with Access-2

System Components. The following are the components of NASA's second MAI system, which is a third generation of the MAI system initiated by Klingbiel at DTIC, and which is accessed interactively online through NASA's electronic Input Processing System (IPS):

- the data file used for MAI, i.e., the Knowledge Base (KB);
- application programs that indicate the input text to be processed, call Access-2, and receive and write the MAI output;
- Access-2, a program that: (1) delineates phrases* found in natural language text by establishing boundaries or parameters to assure grammatically correct word associations without parsing; (2) concatenates words within the established boundaries to form search keys; (3) looks up the search keys in the KB; and (4) returns the candidate NASA Thesaurus posting terms and any other reports to the application program for output to the user. Access-2 was designed to reduce the inefficiencies of the original phrase delineation and logic-code processes (both of which were found to be unnecessary for quality performance of the system) and to shorten the overall processing time.

These components are described in greater detail below.

Data File. The NASA MAI system currently uses a Virtual Storage Access Method (VSAM) file that consists of over 113,000 records (or rules). It consists of two fields:

- Key
- NASA Posting Term(s)

Not only have the number of fields in the KB been reduced, but also the length of each field has been shortened to eliminate the need to read a lot of empty spaces. There is still, however, enough space to accommodate any anticipated input. The key field's null "final word" has been changed from 00 to 999. Otherwise the key and the posting term(s) of each record are entered in the same way as before.

Key: The key is unique and consists of words or phrases that may be encountered in input material. The key field was originally two fields, the initial word and the final word fields. These were concatenated to form a key, which is the address to the record. (See the explanation under Access-1, Data File, Key.) Semicolons

* In delineating phrases, Access-2 uses a stopword list, which, after a number of tests, was stabiliized at about 250 words, down from the more than 76,000 stopwords that were used with MAPS and Access-1. See Appendix C for the current list. The selection of stopwords was based (1) on frequency of occurrence, and (2) comparative lack of indexable concept.

separate the words in a key and every key must contain at least one semicolon. When an entry consists of a single word, the value of the final word field is null. This was previously marked with two zeros, as was done at DTIC, but is now, correctly, marked with nines, because NASA computers sort nines last. Three nines were selected because two nines appear in text frequently enough to cause a problem as a symbol for null. All NASA Thesaurus posting terms and and Use references are included as keys, as well as many additional natural language words and phrases that are the conceptual equivalents of the NASA Thesaurus posting terms.

When the key contains more than two words - i.e., three or more words or two or more words and ";999" - intermediate records, called continuation entries, are created to build to the multi-word key. The key of the first entry consists of the first two words. Each continuation entry adds one more word to the key until the key contains the entire phrase.

Posting Term(s): When the key contains the complete phrase to be translated, the posting term field provides the NASA term or terms that express the same concept. Continuation entries have an asterisk (*) in the posting term field; this tells the computer to look for another word or for ";999" to find a key that will translate. Two zeros in the posting term field mean either that no appropriate NASA translation is available, or that the concept expressed by this key is usually not indexed to by human indexers when the word or words in the key are encountered in natural language text. When a translation is wanted for the key of a continuation entry, a semicolon and three nines are added at the end of the key, to create a unique key. The appropriate translation is then entered in the posting term field.

No logic codes are used with Access-2. Logic codes were previously assigned to each entry in the Knowledge Base. It had been noted, however, that as the KB grew, many non-T logic codes were changed to T's. Approximately half of the words had been found to be context sensitive, and the number was growing. It was decided, therefore, to process all words as if they were context sensitive and had a T logic code, thus saving the input/output (I/O) time required for determining the logic code. This proved to be a good move - for it saved time without adversely affecting the quality of the output - and was subsequently put into production.

Application Programs. The application program differs for each use of MAI. It handles input, output, and interfaces with the Access-2 program. It passes to Access-2 the text from which index terms are to be extracted, receives the MAI selected NASA Thesaurus posting terms, and prints out all required reports.

Access-2. Access-2 was written as a modular program that is called by an applications program. It delineates phrases from input text, constructs search keys from those phrases, looks them up in the Knowledge Base, and returns thesaurus terms to the applications program.

Text Processing with Access-2. Processing of input text by Access-2 is done as follows:

- The computer breaks the input text into words strings by stopping at certain punctuation, such as periods, colons, and semicolons, and any predesignated stopword. (See Appendix C for the Stopword List.)

- These word strings are then examined, from left to right, in five word segments, beginning with word one and word two. The first word of every word combination is checked against the Knowledge Base to see if it exists. If it does not, the word is written out for indexer review.

PROCEDURE:

- If word one followed by word two is found in the Knowledge Base as a key to an entry, the posting term field of that entry, which contains the equivalent NASA Thesaurus term(s), is read. There are three possibilities:

  - The posting term field contains 00, in which case these two words will not generate any posting term.

  - The posting term field contains one or more Thesaurus terms that will be provided to the indexer as suggested indexing terms.

  - The posting term field contains an asterisk. This causes the computer to look for an additional word within the five word segment that, when added to the two previous words, will match the key to another record.

END PROCEDURE.

- If word one followed by word two has an asterisk in the posting term field, and this combination followed by word three, or four, or five does not find a matching key in the Knowledge Base, then the computer adds 999 (which sorts last in the NASA system) in place of the final word, and tries that combination as a key. If that is not found, the final word in the candidate key is dropped, and replaced with 999. This procedure is repeated, if necessary,. until the key is reduced to the first word and 999.

- If word one followed by word two is not found in the Knowledge Base, then word one is concatenated with word three to produce a possible key and PROCEDURE is repeated.

- If word one has been tried with each other word in the five word segment and no key leading to a Thesaurus term is found, the computer looks up word one followed by 999 to see if a Thesaurus term is provided for a single word. This may occur for a strong noun that can stand alone.

- When the process has used or rejected word one, the five word segment is again measured off, beginning with word two.

- Any word in a key that (1) is found in the Knowledge Base and (2) returns an output of 00 or a NASA term (or terms) is poisoned. A poisoned word may not be used in a second key unless an unpoisoned word is added to it.

The following example illustrates a simple case of Access-2 processing for MAI. KB file entries needed to process the sample input, and some related KB entries, have been extracted and are listed below.

Key	Posting Term
ACOUSTIC;DATA	*
ACOUSTIC;DATA;CAPSULE	ACOUSTIC PROPERTIES
ACOUSTIC;DATA;999	ACOUSTIC PROPERTIES
BLADE-VORTEX;INTERACTION	BLADE-VORTEX INTERACTION
BLADE-VORTEX;TURBINE	TURBINE BLADES
BLADE;VORTEX	*
BLADE;VORTEX;INTERACTION	BLADE-VORTEX INTERACTION
BLADE;999	00
BO-105;HELICOPTER	BO-105 HELICOPTER
BO-105;HELICOPTERS	BO-105 HELICOPTER
CLIMB;999	CLIMBING FLIGHT
DATA;999	00
DESCENT;999	DESCENT
HELICOPTER;NOISE	AEROACOUSTICS,AERODYNAMIC NOISE,AIRCRAFT NOISE
HELICOPTER;ROTOR	*
HELICOPTER;ROTOR;NOISE	AEROACOUSTICS,AERODYNAMIC NOISE,AIRCRAFT NOISE
HELICOPTER;ROTOR;999	ROTARY WINGS
HELICOPTER;ROTORS	ROTARY WINGS
TURBULENT;WAKE	TURBULENT WAKES
TURBULENT;WAKES	TURBULENT WAKES
TURBULENT;999	TURBULENCE
WIND;TUNNEL	*
WIND;TUNNEL;TEST	WIND TUNNEL TESTS
WIND;TUNNEL;TESTING	WIND TUNNEL TESTS
WIND;TUNNEL;TESTS	WIND TUNNEL TESTS
WIND;TUNNEL;999	WIND TUNNELS
WIND;TUNNELS;999	WIND TUNNELS

Given the following title and sentences from an abstract of a document:

Helicopter Noise

Acoustic data for a 40 percent model MBB BO-105 helicopter main rotor were obtained from wind tunnel testing and scaled to equivalent actual flyover cases. It is shown that during descent the dominant noise is caused by impulsive blade-vortex interaction (BVI) noise. In level flight and mild climb BVI activity is absent; the dominant noise is caused by blade-turbulent wake interaction.

Phrases are delineated by ending textual word strings whenever a stopword (see Appendix C) or any thought-ending punctuation such as a period, colon, or semicolon is encountered. The phrase delineation process produces the following phrases from this title and abstract:

helicopter noise
acoustic data for a 40 percent model MBB BO-105 helicopter main rotor
from wind tunnel testing and scaled to equivalent actual flyover cases
descent the dominant noise
by impulsive blade-vortex interaction (BVI) noise
in level flight and mild climb BVI activity
absent
the dominant noise
by blade-turbulent wake interaction

In the following Access-2 processing, references are made to the input array and the KB entries shown above.

Processing Description:	Outcome:
Mark off 5-word array in title.	Only 2 words exist, therefore the array is "Helicopter Noise."
Look up search key "HELICOPTER;NOISE" in KB.	Key found. Posting terms "AEROACOUSTICS,AERODYNAMIC NOISE,AIRCRAFT NOISE" returned.
No more words exist in the title.	Move to the first phrase in the abstract.
The first MAI-selected phrase in the abstract is: "Acoustic data for a 40 percent model MBB BO-105 helicopter main rotor."	
PROCEDURE:	
- Mark off the first 5-word array in the phrase. Concatenate word 1 and word 2 to form search key. (In the sample phrase this would be ACOUSTIC;DATA.) Look search key up in the KB.	Key found. Posting term "ACOUSTIC PROPERTIES' returned.
- If the key leads to a posting term(s) or 00, poison (flag) the words in the key (e.g., the words "ACOUSTIC" and "DATA") and end processing for word 1.(e.g., ACOUSTIC).	A poisoned word may not be used again unless it is combined in a search key with an unpoisoned word.

- Move one word to the right in the phrase and mark off a new 5-word array (e.g., DATA FOR A 40 PERCENT). Concatenate the new array's word 1 with the new array's word 2 and look up the search key (e.g., DATA;FOR). Key not found.

- If key is not found, look up the next search search key(s) in this array, that is, concatenate words 1 and 3, 1 and 4, 1 and 5 (e.g., "DATA;A," "DATA;40," "DATA;PERCENT"). Keys not found.

- If an asterisk is found in the posting term field, the key must have an additional word or 999 in order to translate. For example, if words 1 and 3 lead to an asterisk, look next for 1, 3, and 4; 1, 3, and 5; and finally, for 1, 3, and 999.

- Continue processing word 1 whenever a key is not found and untried words remain in the 5-word array.

- End the processing for word 1 whenever the KB provides output for a key, or word 1 has been tried unsuccessfully with words 2, 3, 4, and 5.

END PROCEDURE.

Mark off the next 5-word array in the phrase and repeat the PROCEDURE described. The remaining arrays for the first phrase are:

"FOR A 40 PERCENT MODEL"
"A 40 PERCENT MODEL MBB"
"40 PERCENT MODEL MBB BO-105"
"PERCENT MODEL MBB BO-105 HELICOPTER'
"MODEL MBB BO-105 HELICOPTER MAIN"
"MBB BO-105 HELICOPTER MAIN ROTOR"

Keys not found.

If fewer than five words remain in the phrase, accept smaller segment and follow same procedures. Only four words remain in the sample phrase - "BO-105 HELICOPTER MAIN ROTOR". Mark them off and process key of word 1 and word 2 ("BO-105;HELICOPTER).

Key found. Posting term "BO-105 HELICOPTERS" returned.

Poison (flag) "BO-105" and "HELICOPTER."

These words may not be used again without an unpoisoned word.

End processing for "BO-105."

This was the first word of a key that successfully matched a key in the KB and provided a NASA Thesaurus term.

Three words now remain in the phrase: "HELICOPTER MAIN ROTOR" and this becomes the new array. Look up search key "HELICOPTER;MAIN".

Key not found. (Note: "HELICOPTER" has been poisoned, but it is coupled with "MAIN," which has not been poisoned.)

Look up the next search key: ""HELICOPTER;ROTOR"

Key found. "ROTOR" has not been poisoned. The Posting Term field holds an asterisk (*) which is returned.

An asterisk indicates that another word is needed. There are no more words in the array, therefore, add ";999" to the search key and look up "HELICOPTER;ROTOR;999".

Key found. Posting term "ROTARY WINGS" returned.

Poison (flag) "ROTOR".

Two words now remain in the phrase: "MAIN ROTOR" and this becomes the new array. Look up search key "MAIN ROTOR". "ROTOR" has been poisoned, but "MAIN" has not been part of a key that has been found.

Key not found.

No more words remain in phrase. End processing for this phrase.

Repeat process described for the remaining phrases.

No keys that begin with the first word in the first array are found.

The second five-word array in the next phrase, i.e., "WIND TUNNEL TESTING AND SCALED", illustrates how the KB entries direct the need to concatenate words in an array.

Look up search key "WIND;TUNNEL".

Key found. Posting term field contains an asterisk (*) which requires the addition of another word from the 5-word array.

Add the next word in the array and look up search key "WIND;TUNNEL;TESTING".

Key found. Posting term "WIND TUNNEL TESTS" returned.

The only other search keys found in the above phrases are:

DESCENT;999 (The search keys "DESCENT;THE", "DESCENT;DOMINANT", and "DESCENT;NOISE" were not found, and so the final word was replaced with "999".)	Key found. Posting term "DESCENT" returned.
"BLADE-VORTEX;INTERACTION"	Key found. Posting term "BLADE-VORTEX INTERACTION" returned.
"CLIMB;999"	Key found. Posting term "CLIMBING FLIGHT" returned.
"TURBULENT;WAKE"	Key found. Posting term "TURBULENT WAKES" returned.

In summary, the following terms were suggested:

HELICOPTER NOISE,
AEROACOUSTICS,
AERODYNAMIC NOISE,
AIRCRAFT NOISE,
ACOUSTIC PROPERTIES,
BO-105 HELICOPTERS,
ROTARY WINGS,
WIND TUNNEL TESTS,
DESCENT,
BLADE-VORTEX INTERACTION
CLIMBING FLIGHT, and
TURBULENT WAKES.

Note that the text that was processed contained several hyphenated words. The application program checks each word with an embedded hyphen or virgule (that is, a diagonal line(/)) against the initial word of the keys in the KB. The compound words BO-105 and BLADE-VORTEX would have been found, the hyphens in these words would have been kept, and the compounds would be treated as a single word. However, BLADE-TURBULENT is not found in an initial position in any KB key; therefore the hyphen between these words would be dropped, and the compound would be treated as two words.

KNOWLEDGE BASE BUILDING PROCEDURES

Sources of KBB Entries

There are a number of ways of finding potential new MAI Knowledge Base entries. Of all the ways that we've tried (after entering all NASA Thesaurus terms and official Use references), we think that the statistical Knowledge Base Text Analysis Tool described below is the most efficient. Sources for the NASA KB entries have included the following:

- The NASA Thesaurus terms and Use references.
- Variations of these terms and Use references, including plurals and singluars, logical inversions, and British spellings.
- Words and phrases that frequently occurred in a Key Word in Context (KWIC) listing of:
  - Partially matched or unidentified machine-selected phrases, i.e., phrases that MAI could not completely translate. These were compiled from both operations output and tests of text already in the NASA STI database.
  - Titles from documents in the STI database, formerly available on a microfilm cassette as the KWIC Combined File Index.
- Phrases that were in the DTIC Lexical Dictionary and in scope for NASA.
- Terms defined in NASA SP-7, "The Dictionary of Technical Terms for Aerospace Use," whenever they were synonymous with NASA Thesaurus terms.
- RECON searches of context-sensitive words.
- MAI test run results.
- Consultant evaluations.
- DTIC and DOE thesauri - selected terms.
- Selected phrases identified by the KBB Text Analysis Tool. This was built on a statistical analysis of the NASA STI database's stored titles and abstracts. The use of this tool was begun with the 150 most heavily posted terms, and to date has been used to analyze about 300 terms. Most of the input since June 1988 has been identified through the output of the KB Text Analysis Tool.
- Indexer feedback, which today is the second most used method of finding new KB entries.

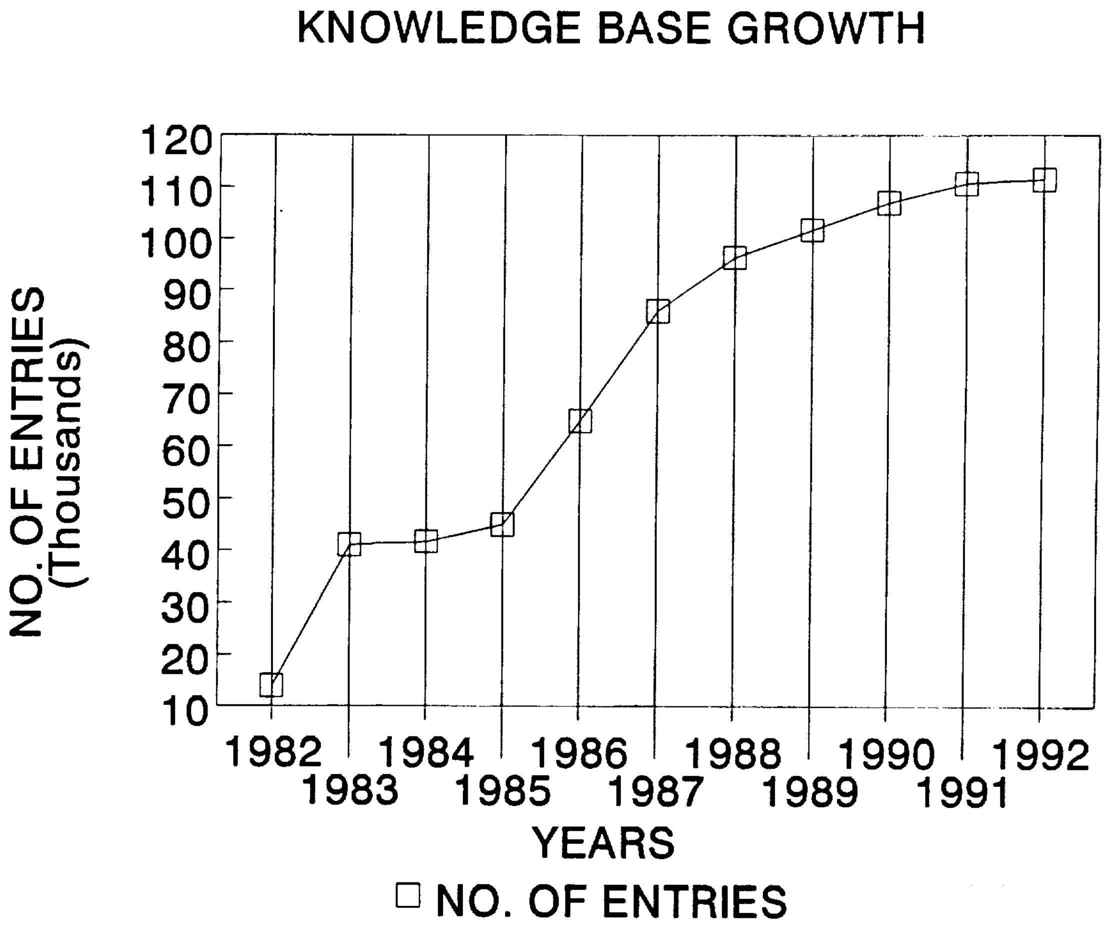
To a large extent, a system can compensate for design tradeoffs by incorporating the appropriate class or classes of entries into the KB. However, in selecting entries for the KB for an online MAI system, the system designers must be concerned with the tradeoffs among the size of the KB file, the system's response time, and the level of complexity selected. By level of complexity is meant the number of special factors to be considered before the system suggests the thesaurus terms for indexer review. These factors might include words containing embedded slashes or hyphens, upper and lower case, the coexistence of broader or narrower suggested terms from a single hierarchy, a category code, and so on. For an online system designer, quick responses have high priority. Regardless of the specific design selected for a machine aided indexing system, its overall performance is largely dependent upon the quality and the comprehensiveness of its Knowledge Base. Strict control and input from domain experts are critical during the database development process. The time and other resources spent in careful construction of the MAI KB pays off with high quality output and indexer acceptance.

Knowledge Base growth is shown in Figure 1. In the period when growth slowed from 1983 to 1984 only one person was assigned to the project and was working on building a Subject Switching database to translate sets of Departments of Energy terms to sets of NASA terms. With the addition of personnel in 1984, the curve begins to rise again. In 1985 the personnel situation was improved further and is reflected in the rather steep climb on the graph. The steady but more gradual slope beginning in June 1988 was caused by the improved quality of terms, made possible by the KBB Text Analysis Tool.

KBB Procedures Used with Access-1

Knowledge Base building during the establishment of the NASA Lexical Dictionary system of MAI was not the fine-tuned process that it is today. The original priority was to make it possible for the computer to recognize every NASA Thesaurus term when it occurred in natural language text. When this was done, the analysts entered all the variants of these terms that came to mind and seemed reasonable. Variants included British spellings. Finally, the analysts entered any synonymous words and phrases that they found in text. When all of this had been done, the primary aids to expanding the KB became lists of phrases that did not completely translate to NASA Thesaurus terms, phrases that could not be identified, and words that could not be found in the NRD. Indexer feedback was strongly encouraged and their suggestions accounted for nearly half of the new entries that were made. In order to find additional material for new KB entries, a program was devised to provide a variety of information about any material run through the NASA Recognition Dictionary (NRD), the Machine Phrase Selection (MAPS) program, and Access-1. We referred to these reports as the "complete package." It included the following information:

1. accession number
2. title
3. abstract
4. major subject terms
5. minor subject terms
6. MAI suggested terms
7. the machine selected phrases
8. "Complete Matches" - phrases that completely matched KB keys
9. "Partial Matches" - phrases that partially matched KB keys
10. "Unidentified Phrases" - phrases that matched no KB keys
11. "Words Not Found in the Recognition Dictionary"
12. "Bad Phrases Sorted by Accession Number", i.e., phrases that were syntactically not able to be handled with existing grammar rules
13. machine selected phrases in three sorts:
    (a) MAPS Phrases Sorted by Accession Number
    (b) MAPS Phrases Sorted by Format Variation, with a summary of format variations
    (c) MAPS Phrases Sorted Alphabetically



YEAR	ENTRIES	YEAR	ENTRIES
1982	14,000	1988	96,300
1983	41,000	1989	101,800
1984	41,700	1990	107,000
1985	45,000	1991	110,700
1986	64,800	1992	111,500
1987	86,000		

1983–1984 shows effect of reduction in resources

FIGURE 1

14. a KWIC (Key Words In Context). Originally this was a KWIC of "good" phrases, i.e., complete and partial matches and unidentified phrases; later it was a KWIC of only those phrases that partially matched.

Items 1 through 7 are illustrated in Figure 2. In this example the analyst has suggested two entries to be made to the KB. Suggestions are reviewed by another analyst or indexer before being entered into the database

Item 8: For an example of Complete Matches, see Figure 3. This printout not only provides an opportunity for the analyst to review existing KB entries for possible improvements, but also it is a source of statistics for evaluating MAI.

Item 9: Phrases with a Partial Match are illustrated in Figure 4. Studying phrases that were only partially translated provided an early approach to identifying new KB entries. This printout also provided statistical and other information for evaluating MAI.

Item 10: For an example of Unidentified Phrases see Figure 5. many unidentified "phrases" are numerics, often measurements. The NASA system usually does not index to numerical concepts, and the original DTIC system excluded numerics from in its controlled vocabulary. Non-numeric unidentified phrases were another source for potential, new KB entries.

Item 11: For a sample list of Words Not Found in the Recognition Dictionary (NRD), see Figure 6. The count to the right of the word indicates how many times the word was encountered in the material processed. Misspelled words on this list are not added to the NRD. Correctly spelled words may or may not be added. (See Appendix A, Category Policy, for a description of the decision-making procedure.) Any word to be added must be researched, categorized, and the transaction entered. Such words might also need to be added to the KB.

Item 12: See the list in Figure 7 for an example of "Bad Phrases." This list provided some insight into word categories, which sometimes needed to be changed. The information gained from this list also caused an occasional change in, or addition to, the MAPS grammar rules. Often this list contained single words that did not express indexable concepts by themselves, but did not appear in text with any clarifying word nearby.

```
                                                  10/ 3/87                PAGE   11

87N26761

Results of Comet Halley observations at the Belgrade Observatory (Yugoslavia)

Comet Halley observations were carried out from Sept. 1985 through June 1986, 75 plates in all being acquired. The resul
ting astrographic positions were communicated to the International Halley watch center. Comet Giacobini-Zinner was also
systematically observed during the period June to Sept. 1985, the observations being intensified in the interval just pr
eceding the ICE spacecraft's closest approach to it. On 9 Dec. 1985 the occultation of BD + 6 deg 5207 by Halley was obs
erved. The results are reviewed.

**** MAJOR SUBJECT TERMS ****

ASTROMETRY                                      SPACE OBSERVATIONS (FROM EARTH)
HALLEY'S COMET

**** MINOR SUBJECT TERMS ****

GIACOBINI-ZINNER COMET                          STELLAR OCCULTATION
*INTERNATIONAL SUN EARTH EXPLORER 3

**** NLD SUGGESTED TERMS ****

GIACOBINI-ZINNER COMET                          INTERNATIONAL COOPERATION
HALLEY'S COMET                                  OBSERVATORIES
[ICE]                                           OCCULTATION

**** PHRASES SELECTED BY MAPS ****

8726761 Z Z H                                  COMET HALLEY OBSERVATIONS

8726761 A J                                    BELGRADE OBSERVATORY

8726761 Z Z H                                  COMET HALLEY OBSERVATIONS

8726761 Z Z                                    75 PLATES

8726761 A Z H Z                                INTERNATIONAL HALLEY WATCH CENTER

8726761 Z Z                                    COMET GIACOBINI-ZINNER

8726761 N                                    T#ICE;SPACECRAFT T#
                                             T# ^ ;oo#ICE
8726761 N                                      OCCULTATION

8726761 Z Z Z P Z                              6 DEG 5207 BY HALLEY
```

Figure 2

```
******** PHRASES WITH A  COMPLETE MATCH ********          DATE:  10/ 3/87                                   PAGE    4
          DENSE;GAS                                        GAS DENSITY

8726757  Z N
  FIBER COLLECTORS
          FIBER                                            FIBERS
          COLLECTORS                                       ACCUMULATORS

8726757  N N
  POLYSTYRENE FOAMS
          POLYSTYRENE                                      POLYSTYRENE
          FOAMS                                            FOAMS

8726757  A Z
  VOLATILE CAPTURE
          VOLATILE                                         OO
          CAPTURE                                          OO

8726757  A A N
  CHEMICALLY ACTIVE GETTERS
          CHEMICALLY                                       OO
          ACTIVE                                           OO
          GETTERS                                          GETTERS

8726757  A N Z Z A Z
  LOW SPEED FLYBY SAMPLE RETURN MISSION
          LOW;SPEED                                        LOW SPEED
          FLYBY;MISSION                                    FLYBY MISSIONS
          SAMPLE;RETURN;MISSION                            SAMPLES

8726758  Z Z
  COMET ATMOSPHERE
          COMET;ATMOSPHERE                                 COMETARY ATMOSPHERES

8726758  A Z
  SLOW FLYBY
          SLOW                                             OO
          FLYBY                                            FLYBY MISSIONS

8726758  Z Z
  10 KM/SEC
          10                                               OO
          KM/SEC                                           OO

8726758  Z Z
  SHORT-PERIOD COMET
          SHORT-PERIOD                                     OO
          COMET                                            COMETS

8726758  N
  COLLECTION
          COLLECTION                                       OO

8726758  Z P Z
  EARTH FOR ANALYSIS
          EARTH                                            OO
          FOR                                              OO
          ANALYSIS                                         OO

8726758  A A A P A H
```

```
********* PHRASES WITH A  PARTIAL MATCH *********          DATE:  10/ 3/87                                    PAGE     1

8726753  A A N
  OTHER WESTERN ARTISTS
         OTHER                                            OO
         WESTERN                                          OO
         ARTISTS                                          **********

8726753  A A P A H P Z
  TEXTUAL VISUAL AND ASTRONOMICAL EVIDENCE TO SUPPORT
         TEXTUAL                                          **********
         VISUAL;AND                                       OO
         VISUAL                                           OO
         AND                                              OO
         ASTRONOMICAL                                     ASTRONOMY
         EVIDENCE                                         OO
         TO:SUPPORT                                       OO

8726753  A A A Z
  GIOTTO PAINTED HIS COMET
         GIOTTO                                           GIOTTO MISSION
         PAINTED                                          **********
         HIS                                              OO
         COMET                                            COMETS

8726753  A A N P Z Z P Z
  REFLECTING HIS VIEWING OF COMET HALLEY IN 1301
         REFLECTING                                       REFLECTION
         HIS                                              OO
         VIEWING                                          VIEWING
         OF                                               OO
         COMET;HALLEY                                     HALLEY'S COMET
         IN                                               OO
         1301                                             **********

8726754  A Z Z Z H
  NASA C-141/KAO COMET HALLEY OBSERVATIONS
         NASA                                             OO
         C-141/KAO                                        **********
         COMET;HALLEY                                     HALLEY'S COMET
         OBSERVATIONS                                     OO

8726754  A Z J
  LEAR JET OBSERVATORY
         LEAR;JET                                         **********
         JET                                              OO
         OBSERVATORY                                      OBSERVATORIES

8726754  A Z A P Z P Z Z
  SPECTRAL RANGE EXTENDED FROM 2 TO 800 MICRONS
         SPECTRAL                                         SPECTRA
         RANGE                                            OO
         EXTENDED                                         OO
         FROM                                             OO
         2                                                OO
         TO                                               OO
         800                                              **********
         MICRONS                                          OO
```

Figure 4

```
******** UNIDENTIFIED PHRASES ********                    DATE:  10/ 3/87                              PAGE     1

8726753  N
  ARTISTS

8726771  Z Z
  NEAR-NUCLEUS TV-IMAGES

8726778  Z Z Z
  0.25 PARTICLE/SQ M/SEC

8726778  Z Z
  53,000/SQ M/SEC

8726784  Z Z Z
  12 POWER/SQ M/SEC

8726786  Z H
  0740 UT

8726786  Z H
  1200 UT

8726786  Z Z
  3H 40M

8726799  N
  WAVEFRONT

8726805  Z Z
  2-O C2

8726815  N
  EIGENFREQUENCIES

8726827  N
  MICROOPTICS

8726851  N
  OPTIMISM


**** TOTAL UNIDENTIFIED PHRASES =      13


**** TOTAL RECORDS READ =    1598
```

```
***  WORDS NOT FOUND IN THE RECOGNITION DICTIONARY  ***            DATE:  10/ 3/87                    PAGE     1

                                                     6
          8726813   8726818   8726821   8726822   8726823   8726824
   ANTICORRELATE                                     1
          8726824
   AQUARIDS                                          1
          8726784
   ARCMIN                                            2
          8726788   8726789
   ASSYMMETRIES                                      1
          8726795
   BLUEWARDS                                         1
          8726813
   BONDONE                                           2
          8726753   8726753
   BVJK                                              1
          8726808
   COMENIUS                                          1
          8726764
   COMETOPAUSE                                       2
          8726795   8726796
   CYANOJETS                                         1
          8726781
   ECLIPTICAL                                        1
          8726764
   ENDROFF                                           1
          8726820
   EURECA                                            3
          8726841   8726841   8726841
   FRACTAL                                           1
          8726811
   GIACOBINID                                        1
          8726786
   HABIG                                             1
          8726830
   HEHAVIOR                                          1
          8726848
   HEISSLER                                          1
          8726812
   ICQ                                               4
          8726759   8726759   8726759   8726759
   INCONTROVERTIBLE                                  1
          8726755
   JHK                                               2
          8726783   8726783
   JOINTLESS                                         2
          8726837   8726837
   LJO                                               2
          8726754   8726754
   LJOS                                              1
          8726754
   MEDVEDEV                                          1
          8726767
   MILLENIUM                                         1
          8726816
   NIEDNER                                           1
          8726794
   OPTOCENTER                                        1
          8726807
```

Figure 6

```
8726769          EARLY                                         A
  40 A
  EARLY

8726769          A                                             G
  40 G
  A

8726769          A                                             G
  40 G
  A

8726769          PHYSICAL                                      A
  40 A
  PHYSICAL

8726769          FAR                                           A
  40 A 41 A
  FAR ONLY

8726770          DIFFERENCE                                    A
  40 A 41 P
  DIFFERENCE IN

8726770          INSTRUMENT                                    Z
  40 Z 41 P
  INSTRUMENT OR

8726770          ONE                                           A
  40 A 41 P
  ONE ON

8726770          HALLEY                                        Z
  40 Z
  HALLEY

8726770          1910                                          Z
  40 Z
  1910

8726770          ANALYSIS                                      Z
  40 Z 41 P
  ANALYSIS OF

8726770          COMPENSATING                                  Z
  40 Z 41 P
  COMPENSATING FOR

8726770          A                                             G
  40 G
  A

8726771          VEGA-SPACECRAFT                               Z
  40 Z
  VEGA-SPACECRAFT
```

Item 13: The three sorts of Machine Selected Phrases are illustrated in Figures 8, 9, and 11. Figure 8 shows MAPS phrases sorted by accession number, which was useful for evaluating how well MAPS performed for individual documents. It also helped to identify phrases that could be researched and possibly added to the KB. The second sort, MAPS Phrases Sorted by Format Variations, served not only as a check on the grammar rules and how well they were performing, but also identified many indexable concepts, and counted the number of times each format occurred in the body of text being analyzed. Figure 9 shows an example of some MAPS phrases sorted by format variations, illustrating the A N (i.e., adjective, strong noun) format and several others. A sample of the report on the count, i.e., a "Summary of Format Variations", is shown in Figure 10. The formats are sorted alphabetically. For a body of 150 documents tested, the most frequently occurring format was the single word/strong noun term, with 93 occurrences. This proved to be a problem and was addressed later. See the section on the Single-Word Term Assessment Tool. A portion of the summary, with examples of equivalent phrases, and sorted by frequency is provided below:

Number of Occurrences	Format	Example
93	N	feasibility
67	ZZ	flight test
54	AZ	lunar surface
41	AN	spanwise blowing
31	ZH	computer codes
30	ZN	helicopter airframes
25	ZZZ	aircraft power systems
19	AZZ	subsonic transport aircraft
17	AH	electron beams
16	AZN	cruise fuel consumption
15	AAZ	forward swept wing
13	ZZH	wind tunnel tests
11	NZ	heat transfer
10	ANZ	UV light sources
10	AZH	high angle-of-attack configurations
10	ZPN	dynamics of loading

```
ACCNUM            FORMAT                                    PHRASE
8470018  Z P A N                      PROGRAM IN POLYMERIC COMPOSITES
8470018  N                            RESEARCH
8470018  Z P T Z N P Z Z              PROPERTIES OF THE CONSTITUENT FIBERS AND MATRIX PROPERTIES
8470018  A A N N                      ADVANCED STRUCTURAL ANALYSIS METHODS
8470018  Z Z P N                      FATIGUE RESPONSE OF LAMINATES
8470018  A Z Z P Z Z P N Z A H        ENVIRONMENTAL RESPONSE MODELING AND PROCESSING SCIENCE FOR LIGHT WEIGHT
                                        AIRFRAME STRUCTURES
8470018  Z P A Z                      PROJECT IN CERAMIC MATERIALS
8470018  A N P Z Z P Z Z              CRITICAL RESEARCH IN MATERIAL PERFORMANCE AND DESIGN METHODOLOGY
8470018  A Z                          BRITTLE MATERIALS
8470018  Z P Z                        PROCESSING AND PROPERTIES
8470018  N P N Z P N N Z              CHARACTERIZATION OF SILICON NITRIDE AND SILICON CARBIDE MATERIALS
8470018  A Z Z Z P N N P A N A Z P Z Z ENVIRONMENTAL RESPONSE PROCESSING SCIENCE AND IMPACT BEHAVIOR OF HIGH
                                        TEMPERATURE CERAMIC BODIES FOR GAS TURBINE ENGINE APPLICATIO
8470018  A N P A A P Z P A A Z        ADVISORY SERVICES TO GUIDE R AND D IN ADVANCED AEROSPACE MATERIALS
8470018  A Z A Z                      NATIONAL MATERIALS ADVISORY BOARD
8470018  Z P Z                        SCIENCE AND ENGINEERING
8470019  A N Z                        HIGH TEMPERATURE MATERIALS
8470019  Z P Z P Z N P Z A P A Z      LEVEL OF MATERIALS AND PROCESSING TECHNOLOGIES FOR HIGH-TEMPERATURE
                                        METALLIC AND CERAMIC MATERIALS
8470019  N                            RELIABILITY
8470019  A N                          STRUCTURAL EFFICIENCY
8470019  Z N                          TURBINE ENGINES
8470019  N                            WORK
8470019  N P Z Z                      INTERRELATIONSHIPS BETWEEN MATERIAL COMPOSITION/MICROSTRUCTURE
8470019  N Z                          FABRICATION PROCESSES
8470019  Z Z                          MECHANICAL/PHYSICAL PROPERTIES
8470019  N P A Z H P N                CREATION OF ADVANCED MATERIALS CONCEPTS AND OPTIONS
8470019  A Z Z Z Z N Z H              HIGHER PERFORMANCE/HIGHER DURABILITY/LOWER COST AIRCRAFT PROPULSION SYSTEM
                                        COMPONENTS
8470019  N                            RESEARCH
8470019  N N                          BASIC STUDIES
8470019  A P Z P Z                    INFLUENCE ON MICROSTRUCTURE/PROPERTIES OF REDUCTIONS
8470019  Z H P A N P N                SUBSTITUTE ELEMENTS FOR CRITICAL METALS IN SUPERALLOYS
8470019  Z Z H P Z Z P N              POTENTIAL IRON-BASE ALLOY OR ALUMINIDE REPLACEMENTS FOR SUPERALLOYS
8470019  A P N N P Z N                SUPPLEMENTED BY BASIC RESEARCH ON CERAMICS/CERAMIC COMPOSITES
8470019  N N H                        BASIC STUDIES FOCUS
8470019  A Z P A N H H P Z N P N      PHASE COMPOSITION/DISTRIBUTION AND ADVANCED FABRICATION PROCESS VARIABLES
                                        FOR CAST/WROUGHT/POWDER METALS AND CERAMICS
8470019  A N Z                        RAPID SOLIDIFICATION TECHNOLOGY
8470019  N Z                          MELTING SPINNING
8470019  N P Z Z Z Z                  STUDIES OF POTENTIAL SERVICE ENVIRONMENT ATTACK
8470019  Z Z                          OXIDATION/HOT CORROSION/ETC
8470019  A P A Z N P A P Z N P A N    CONTROLLED AND SIMULATED ENGINE ENVIRONMENTS TO GUIDE AND SUPPORT BASIC AND
                                        APPLIED RESEARCH
8470019  N P A A P A Z N H            VALIDATION OF ADVANCED METALLIC AND THERMAL BARRIER COATING CONCEPTS
8470019  N N                          TRIBOLOGY RESEARCH
8470019  Z H S                        MATERIAL/LUBRICATION/WEAR INTERACTION
8470020  A A N                        ADVANCED STRUCTURAL ALLOYS
8470020  N                            RESEARCH
8470020  A Z H Z Z P A A N            METALLURGICAL STRUCTURE/MECHANICAL PROPERTY RELATIONSHIPS CHARACTERISTIC OF
                                        ADVANCED STRUCTURAL ALLOYS
8470020  A P A H                      NEW OR IMPROVED CONCEPTS
8470020  A N                          STRUCTURAL ALLOYS
8470020  Z Z                          AIRCRAFT APPLICATIONS
8470020  Z N                          CURRENT RESEARCH
8470020  N P T Z Z                    STUDIES OF THE STRUCTURE/PROPERTY RELATIONSHIPS
```

MAPS PHRASES SORTED BY FORMAT VARIATIONS

```
 8470360  AHPZNPN   DIRECT MEASUREMENTS OF RATE CONSTANTS AND TEMPERATURE
   TOTAL  AHPZNPN   1

 8470356  AHZ       WAVELENGTH REGION 180-300
   TOTAL  AHZ       1

 8470365  AHZH      LONG PATH GAS CELLS
   TOTAL  AHZH      1

 8470356  AN        HIGH ALTITUDE
 8470356  AN        OPTICAL DEPTH
 8470360  AN        CHEMICAL KINETICS
 8470361  AN        MARINE ALGAE
 8470361  AN        MARINE ALGAE
 8470363  AN        ATMOSPHERIC PHOTOCHEMISTRY
 8470363  AN        POLAR MOLECULES
 8470364  AN        ATMOSPHERIC PHOTOCHEMISTRY
 8470365  AN        SPECTROSCOPIC TECHNIQUES
 8470365  AN        MOLECULAR PARAMETERS
 8470366  AN        SPECTRAL RESOLUTION
 8470367  AN        SPECTRAL PARAMETERS
 8470370  AN        IDENTIFY GAPS
 8470371  AN        PLANETARY GEOLOGY
 8470371  AN        TERRESTRIAL SOILS
 8470371  AN        VOLCANIC STRATIGRAPHY
 8470373  AN        PLANETARY GEOLOGY
 8470374  AN        MARTIAN VOLCANOES
 8470376  AN        COSMIC DUST
 8470377  AN        ROCK SAMPLES
 8470377  AN        THERMODYNAMIC PARAMETERS
 8470377  AN        ROCK SAMPLES
 8470377  AN        COSMIC DUST
 8470379  AN        ABSOLUTE TIME
 8470379  AN        ROCK SAMPLES
 8470379  AN        RADIOACTIVE NUCLIDES
 8470380  AN        NOBLE GASES
 8..etc.  AN        .....
   TOTAL  AN        56

 8470798  ANAZZZN   HIGH RESOLUTION GROUND BASED % MM IMAGES
   TOTAL  ANAZZZN   1

 8470351  ANH       INERTIAL NAVIGATION UNIT
 8470357  ANH       REMOTE SENSING INSTRUMENTS
 8470361  ANH       ATMOSPHERIC OZONE LAYER
 8470379  ANH       LARGE IMPACT CRATER
 8470382  ANH       PRIMITIVE ENSTATITE CHONDRITE
 8470382  ANH       NUCLEAR RESONANCE TECHNIQUE
 8470387  ANH       SECULAR RESONANCE SURFACES
 8470393  ANH       ATMOSPHERIC TEMPERATURE STRUCTURE
 8470398  ANH       COMETARY DUST STUDY
   TOTAL  ANH       9
```

Figure 9

```
    10/15/84              SUMMARY OF THE FORMAT VARIATIONS                                    PAGE:     1

TOTAL A A A Z H P T Z P N P N N       1
TOTAL A A H                           6
TOTAL A A H H                         1
TOTAL A A H P Z N                     1
TOTAL A A H P Z P Z                   1
TOTAL A A H Z                         1
TOTAL A A N                           5
TOTAL A A N H                         1
TOTAL A A N N                         2
TOTAL A A N P A H                     1
TOTAL A A N P N                       1
TOTAL A A N P T N P Z P N P N         1
TOTAL A A N Z                         1
TOTAL A A P A N                       1
TOTAL A A P A N S                     1
TOTAL A A P A Z                       1
TOTAL A A P N N                       1
TOTAL A A P Z                         1
TOTAL A A Z                          12
TOTAL A A Z A N P N H                 1
TOTAL A A Z H P A Z A Z               1
TOTAL A A Z N P Z Z N H B A P A A A   1
TOTAL A A Z P A Z P A N Z             1
TOTAL A A Z P N                       1
TOTAL A A Z Z                         2
TOTAL A H                            22
TOTAL A H A A P A H                   1
TOTAL A H N                           1
TOTAL A H P A A Z P N N P T Z         1
TOTAL A H P A N N                     1
TOTAL A H P A Z P Z N Z               1
TOTAL A H P H P A P A N               1
TOTAL A H P T Z                       1
TOTAL A H P T Z Z                     1
TOTAL A H P Z                         1
TOTAL A H P Z A H P A N P T N A N     1
TOTAL A H P Z N                       1
TOTAL A H P Z N P H                   1
TOTAL A H P Z Z H                     1
TOTAL A H P Z Z N                     1
TOTAL A H Z Z N                       1
TOTAL A N                            61
TOTAL A N A H                         1
TOTAL A N A N                         1
TOTAL A N A N Z H                     1
TOTAL A N H                           6
TOTAL A N H Z                         1
TOTAL A N N                           6
TOTAL A N N H                         1
TOTAL A N N H Z                       1
TOTAL A N N N                         3
TOTAL A N N P T Z H                   1
TOTAL A N P A A P Z P A A Z           1
TOTAL A N P A N                       1
TOTAL A N P A P A Z P Z               1
TOTAL A N P A Z                       2
TOTAL A N P H                         1
```

The third sort of MAPS phrases, an alphabetical list of the MAPS phrases, is shown in Figure 11. This printout provided the analyst with some information on the frequency of occurrence of any given phrase in that body of text.

Item 14, a Key Words in Context (KWIC) listing of MAPS phrases that partially matched KB keys, is illustrated in Figure 12. The NLD team's use of KWIC indexes for identifying repetitious phrases was begun in August 1985. While this kind of index was not new, this innovative use of KWICs allowed the NLD team to double the production of new entries made to the KB in the first month. The method was definitely an improvement over former methods, but it also was frustrating when no equivalent NASA term(s) could be found for an often-repeated phrase. Conversely, such phrases were helpful in that they pointed out deficiencies in the NASA Thesaurus. Suggestions for new thesaurus terms to eliminate these deficiencies were passed along to the Lexicographer for addition to the controlled vocabulary.

In a text-based MAI system such as NASA's, semantic analysis is dependent upon the content of the KB. Primary concerns with this functional element were the slow development time, the level of manual effort associated with KB construction and maintenance, the need to generate high quality output, and the problem arising from limiting analysis to certain syntactic phrase forms.

KBB Procedures Used with Access-2

With the change from Access-1 to Access-2 there has also been a change in the procedures used for building the Knowledge Base (KB).

The KB used with Access-2 has a relatively simple structure. As described earlier, it now contains two fields:

- o the Key field, which holds the natural language input word or phrase that is also the address to the record in the computer file; and

- o the Posting Term field, which contains the semantically equivalent NASA Thesaurus term(s).

Variant forms of a word are included in the KB keys because NASA MAI experience indicated that word-stemming tends to result in ambiguous word forms and cause unwanted output. Words containing hyphens are searched once with the hyphen, and, if not found, are then searched without the hyphen. Ambiguous terms in the text may be disambiguated by the choice of entries in the KB, or ambiguous terms may be left out of the KB or receive a null translation, that is, be posted to two zeros (00). If the content of the search key (semantic unit) is not of indexable importance, the KB can delete the unit either by omitting the unit from the KB or by providing a posting of 00. Recognition of the semantic value of each unit depends upon the content of the KB, which, in turn, depends upon the evaluation of entries made by the analysts who created the KB and continue to add records to it.

ACCNUM	FORMAT	PHRASE
8726823	A H	TWO FLUIDS
8726817	A Z N	TWO LONG-TAILED GALAXIES
8726767	A Z P A H	TWO MODELS OF NONGRAVITATIONAL FORCES
8726821	A H	TWO PICTURES
8726784	A N P N	TWO POPULATIONS OF PARTICLES
8726800	A Z Z	TWO WAVE MODES
8726823	Z A Z	TWO-FLUID HYDRODYNAMICAL MODEL
8726839	A P Z Z Z	TYPE OF ANTENNA TEST FACILITY
8726759	A P Z P N	TYPE OF DATA FOR COMETS
8726764	Z J	UNIVERSITY OBSERVATORY
8726825	A H	UPPER LIMIT
8726799	N H	UPSTREAM WAVES
8726799	N H P A Z	UPSTREAM WAVES OF COMETARY ORIGIN
8726807	A N	USABLE PHOTOMETRY
8726793	A A Z Z	V X B PICKUP
8726836	A N	VARIABLE CAMBER
8726815	A H	VARIATIONAL CALCULATIONS
8726781	N P A Z	VARIATIONS IN RELATIVE ABUNDANCE
8726788	N P A P Z Z P N	VARIATIONS OF CENTRAL AND INTEGRAL MAGNITUDES AND COLORS
8726833	Z Z Z	VECTOR COMPUTER PERFORMANCE
8726803	A Z H	VEGA AVP-V EXPERIMENT
8726794	A N	VEGA ENCOUNTERS
8726794	A Z	VEGA SPACECRAFT
8726792	A Z N	VEGA SPACECRAFT ENCOUNTERS
8726774	A Z P Z Z Z N Z Z Z Z	VEGA 1 AND 2 COMET HALLEY ENCOUNTERS PLASMA WAVE INSTRUMENT APV-N
8726775	Z P Z	VEGA-1 AND VEGA-2
8726794	Z P Z	VEGA-1 AND VEGA-2
8726803	Z P Z	VEGA-1 AND VEGA-2
8726796	Z P Z	VEGA-1 AND 2
8726762	N	VELOCITY
8726776	N	VELOCITY
8726791	N	VELOCITY
8726802	N	VELOCITY
8726847	N P N	VELOCITY OF DETONATION
8726824	N Z	VELOCITY SELECTION
8726782	N	VENUS
8726790	A A Z	VERY CLOSE APPROACHES
8726825	A A Z P A N	VERY HIGH DENSITY OF IONIZING PROTONS
8726845	A A Z Z Z	VERY HIGH ENERGY LASER SYSTEMS
8726800	A A Z A H	VERY LOW FREQUENCY ELECTROSTATIC EMISSIONS
8726846	A A Z	VIDEO REAL TIME
8726782	A P A Z P A Z Z P A Z Z	VISIBLE IN BROAD DAYLIGHT AT THETA 1 DEG AND THETA 42 DEG
8726754	A N	VISIBLE PHOTOGRAPHY
8726782	Z N	VISIBLE-TO-INFRARED EMISSION
8726759	A N	VISUAL ESTIMATES
8726769	A N N P A Z	VISUAL MAGNITUDE ESTIMATES OF HALLEYS COMET
8726763	A Z	VISUAL OBSERVATION
8726763	A Z P Z Z Z	VISUAL OBSERVATION OF COMET HALLEY 1985-86
8726763	A Z P A Z	VISUAL OBSERVATION OF HALLEYS COMET
8726770	A Z	VISUAL OBSERVERS
8726819	N	VOIDS
8726757	A Z	VOLATILE CAPTURE
8726781	A A H	VOLATILE ICY GRAINS
8726755	A Z	VOLATILE SPECIES
8726821	A A P H N P Z N Z	VOYAGER I AND II IMAGES OF JUPITERS NIGHT HEMISPHERE
8726821	A N P A N	VOYAGER IMAGES OF JOVIAN LIGHTNING
8726798	Z A P Z Z	VO SIMILAR TO 60 KM/SEC

```
10/ 3/87                                        KWIC OF PARTIAL PHRASES FROM STAR                                          PAGE   13

8726825                                            LINE-EMITTING CLOUDS
8726817                                            LINES OF INQUIRY
8726844                                            LOCATION BY DEFENSIVE ARMS
8726801                                            LONG PERIOD WAVES
8726817 TWO                                        LONG-TAILED GALAXIES
8726756                                            LOW COST OR HOME
8726844 HIGH EFFICIENCY AND                        LOW COSTS
8726795                                            LOW ENERGY ELECTRONS AND SPACECRAFT POTENTIALS NEAR COMET HALLEY
8726785                                            LOW INCLINATIONS AND SHORT REVOLUTION PERIODS
8726789                                            LOW RESOLUTION MAPPING OF COMET HALLEY IN PRINCIPAL ATOMIC AND MOLECULAR SPECIES
8726825 SPECTROSCOPIC DATA ON 12                   LOW-REDSHIFT
8726818 SIXTY-THREE                                LOW-RESOLUTION IUE SPECTRA OF 124 WELL-CLASSIFIED O3 TO B5 STARS
8726818                                            LOW-RESOLUTION IUE SPECTRA
8726840 PAYLOAD CAPACITY OF 10T FOR                LOWER GEOSYNCHRONOUS ORBITS
8726799 SIMILAR OR                                 LOWER LEVELS
8726817                                            LUMINOUS MERGING GALAXY
8726817                                            LUMINOUS PECULIAR GALAXIES M82
8726821                                            LUMINOUS SPOTS
8726822 1.8                                        M FOCAL LENGTH EBERT-FASTIE MONOCHROMATOR
8726802                                            M 1.5
8726784 12 POWER/SQ                                M/SEC
8726778 0.25 PARTICLE/SQ                           M/SEC
8726778 IMPACT RATE AVERAGES 5000/SQ               M/SEC
8726778 53,000/SQ                                  M/SEC
8726830 INTAKE MODELS FOR                          MACH NUMBERS FROM 0 TO 1.9
8726822 STANDARD PARAMETERS OF SOLAR               MAGNETIC ACTIVITY
8726794 SECTOR BOUNDARY/FRONTSIDE                  MAGNETIC RECONNECTION MODEL
8726814 THREE-DIMENSIONAL                          MAGNETOHYDROSTATIC EQUILIBRIA
8726801                                            MAGNETOSONIC MODE
8726814 SHORT-LIVED                                MAGNETOTAIL
8726837 BEARINGLESS                                MAIN ROTOR SYSTEM
8726789 LOW RESOLUTION                             MAPPING OF COMET HALLEY IN PRINCIPAL ATOMIC AND MOLECULAR SPECIES
8726849 SEMIAUTOMATIC COMPARISON OF                MAPS AND RECONNAISSANCE VIDEO DATA
8726850                                            MARKOV MODELS
8726767                                            MARSDEN-SEKANINAS MODEL
8726784 DIFFERENTIAL                               MASS INDEX S 1.85
8726798                                            MASS-LOADED REGION
8726818                                            MASS/SPECTRAL TYPE CLASS RELATIONSHIPS
8726792 VARIATIONS IN DUST COUNT RATES FOR         MASSES
8726769                                            MATHEMATICAL FORMULAE
8726830 DYNAMIC                                    MEASUREMENTS FROM 3
8726794 SOLAR WIND/IMF                             MEASUREMENTS
8726844 IN DEFENSIVE MINES AND STEERING            MECHANISMS
8726817 LUMINOUS                                   MERGING GALAXY
8726785 HISTORICAL RECORDS OF COMETS AND           METEOR SHOWERS IN CHRONICLES
8726784 HALLEY                                     METEOR SHOWERS IN 1985-1986
8726785 COMETS AND                                 METEOR SHOWERS OF 461
8726785 BETWEEN ANCIENT COMETS AND                 METEOR SHOWERS
8726830 LAMBDA/2                                   METHOD
8726842 CONTROL SYSTEMS AND                        METHODS FOR CONTROL SYNTHESIS AND OPTIMIZATION
8726780 HIGHER RATIOS                              MG/SI,FE/SI,AL/SI IN COMPARISON
8726822 2800                                       MHZ EMISSIONS
8726786 SCATTERING OF RADIO WAVES AT 70.31         MHZ
8726804 4.6 TO 10.3                                MICRON COLOR TEMPERATURE DEPENDENCE ON HELIOCENTRIC DISTANCE
8726817 MULTICOLOR 1 TO 3                          MICRON GALAXY SCANS
8726776 2.5                                        MICRON THICK ALUMINIZED MYLAR FILM
8726804 10.3                                       MICRONS
8726804 OF COMET HALLEY AT 2-13                    MICRONS
```

Figure 12

Several methods for identifying the various expressions that are synonymous with thesaurus terms are described below. This process may seem like an infinite task, especially if the thesaurus is large - and time consumption is not the only drawback to the first three methods described. All can lead to an unnecessarily large Knowledge Base due to the addition of expressions that are essentially either unique or have a very low frequency of occurrence. Obviously text words that are not anticipated and included in the KB cannot be translated to the controlled vocabulary (Artandi, 1976). However, the probability of this occurring can be greatly reduced by using the latest KBB procedure that NASA developed after trying the others.

o A review of abstract text on a case-by-case basis was scarcely considered inasmuch as it is highly inefficient.

o A review of partially matched and unidentified MAPS phrases in the various sorts already described in Item 13 above and illustrated in Figures 8, 9, and 11, was helpful and produced new KB entries, but it was untargeted with regard to domain concepts.

o KWIC (Key Words In Context) indexes of available text, while an improvement over earlier methods of identifying text phrases, are of limited use since they too are untargeted. Their use was often frustrating precisely because, as arrangements by "key words", these lists could leave you with indexable concepts for which no thesaurus terms could be substituted.

o Lists of "Words Not Found As the First Element of a KB Key", see Figure 13. Any frequently-used, correctly-spelled word is a candidate for addition to the KB. Words that do not express an indexable concept either alone or in combination with other words are added to the KB with a null posting to improve the IPS spell-checking capabilities. (See a description of this use in the Introduction under Scope.) The list also alerts the MAI staff to new words and phrases that may be coming into use to describe new technology. These new concepts may require the addition of new terms to the NASA Thesaurus.

o The NASA-developed Knowledge Base Building (KBB) Text Analysis tool.

```
*** WORDS NOT FOUND IN NLD AS FIRST ELEMENT OF KEY ***          DATE:  1/17/91                    PAGE:   4

  1  MANUFACTURER-PROVIDED         115  8835821
  2  MASS-PRODUCIBLE               119  8519543   119  8519543
  1  MASTER                        119  8519543
  1  MEASURMENTS                   119  8724858
  1  MECHAISM                      119  8129008
  1  MERIDIONALLY                  119  9014730
  1  MEUDON                        119  8724386
  2  MICHELS                       119  9029328   119  9029328
  1  MID-INFRARED                  137  8573724
  1  MIDCHORD                      119  8722181
  1  MISR                          119  8519543
  2  MOIST                         119  8821387   119  9110516
  1  MONOCHROME                    115  8042844
  3  MPM                           119  8821387   119  8821387
  1  MULTI-STURCTURED              119  9024560
  2  MULTICOLOR                    115  8042844   115  8042844
  1  MULTILINEAR                   119  8825089
  1  MULTIWINDOW                   119  9018394
  1  MUNCH                         119  8915500
  1  N-TYPE                        119  7933085
  1  NARROWS                       119  9024560
  2  NATIVE                        119  8129008   119  8129008
  5  NATURAL-LAMINAR-FLOW          119  9014210   119  9029328
  1  NEAR-                         137  8573724
  1  NEAR-WALL                     119  9024560
  1  NEGEV/NORTHERN                119  8811231
  3  NEMATIC                       119  8917013   137  8470452
  1  NEUTRALS                      139  9070371
  1  NEWLY-DEVELOPED               117  8810425
  1  NICHOLLS                      119  9028262
  1  NICOLLS                       119  9028262
  1  NITROGEN/WATER                117  8810171
  2  NON-CONTIGUOUS                119  8826712   119  8826716
  1  NON-REACTIVE                  139  9070371
  1  NONDIMENSIONAL                119  8918664
  1  NORTH-SOUTH                   119  9110516
  1  NOSETIP                       136  8772181
  1  NOVAYA                        115  9048391
  1  OIL-FILM                      115  8835821
  1  ONE-FLUID                     115  8845047
  1  ONE-FOURTH                    119  8629134
  1  OPEN-LITERATURE               119  8826740
  1  OPTO-ELECTRONICAL             137  8470452
  1  OR/AND                        119  8715424
  1  OUGHT                         119  8819824
  3  OVERGRAZING                   119  8811231   119  8811231
  1  OVERHEAD                      119  8827677
  2  OVERLAP                       119  8819824   119  8917013
  1  OVERSHOOTS                    136  9072383
  1  P-ANISALDAZINE                119  6921227
  1  P-TYPE                        119  7933085
  1  PACK                          119  8917013
  1  PARTICLE-GAS                  119  8815747
  1  PARTICULARITY                 119  8517705
  1  PC/CRT/FILM                   119  7316228
  1  PC/CRT/GLASS                  119  7316228
  1  PC/LASER/FILM                 119  7316228
  1  PC/LASER/GLASS                119  7316228
```

Figure 13

KBB Text Analysis tool.

The NASA-developed KBB program is a statistically based text analysis tool that presents the domain expert with a well-filtered list of synonymous and conceptually-related phrases for each thesaurus concept. This tool was designed to satisfy three main requirements:

1. The output phrases for any given use of this tool would be targeted to one specific thesaurus concept - thus all expressions related to a particular target term could be analyzed together. By targeting, in separate operations, all the terms in a single hierarchy, or all of the terms that share a word (such as MATRICES, MATRICES (CIRCUITS), and MATRICES (MATHEMATICS)), expressions that can lead to ambiguity can be identified and analyzed as well.

2. The output phrases would be restricted to those that had a high frequency of occurrence within the existing NASA database - thus screening out "unique" expressions.

3. The phrases would be normalized, i.e., of the same structure as phrases extracted by the semantic-unit identification operation.

The basic processing steps of the KBB Text Analysis Tool illustrated in Figure 14 can be described as follows:

- Input text is selected. Generally, this will be the titles and abstracts of a large set of document records (150-1,000) indexed to, or otherwise identified as being related to, a single thesaurus concept. At NASA, a standard online RECON search is used to identify an accurate set of such records.
- The text is copied into a file and preprocessed using a simple text breaking method similar to that used for delineating text phrases for MAI, described earlier under the example of Access-2 processing.

- A concatenation process is then used to identify all possible multiword phrases within a maximum length along with certain rules that provide syntactic filtering (which, for example, prevent prepositions and articles from beginning or ending a phrase).

- A count of the frequency-of-occurrence is determined for each unique single-word and multi-word phrase. Then the words and phrases are sorted in descending order by the frequency values. A lower-limit value is established and phrases with fewer occurrences than that value are eliminated. There is a natural bias for single-word phrases to have much higher frequencies than two-word units, which in turn, will have higher frequencies than three-word phrases, etc. This can be dealt with in two ways. A simple way is to produce five separate sorts, each one corresponding to a different phrase length. The other is to use a derived frequency value that effectively accounts for the bias. A process for determining such a value was recently described by Jones, Gassie, and Radhakrishnan (1990). The formula can be stated as W*F*N2 where W is the sum of the frequencies of the words in the phrase, F equals the frequency of the phrase, N2 equals the number of distinct words in a phrase, squared, and the asterisks indicate multiplication.

KNOWLEDGE BASE BUILDING TOOL

INPUT	PROCESSING	OUTPUT
TEXT FIELDS FROM A CONCEPT-SPECIFIC SEARCH OF THE NASA STI DATABASE	TEXT-BREAKING ROUTINE WORD CONCATENATION PROCESS FREQUENCY SORT PHRASE FILTERING (KB LOOK-UP, NORMALIZATION)	WORD AND PHRASE LISTS CONTAINING FREQUENTLY OCCURRING NATURAL LANGUAGE 'SYNONYMS'

Figure 14

- The final processing procedure further refines the output. The phrases are checked against the existing KB entries to eliminate (1) any phrase that properly translates to a thesaurus concept other than the one that the KBB is currently analyzing; and (2) single words or multi-word phrases that have been identified as having a poor or low semantic value.

Sample output from the Knowledge Base Building Tool (KBB) is shown in Figure 15. The input consisted of titles and abstracts from records associated with the thesaurus concept METAL MATRIX COMPOSITES. The first column in this figure lists the unedited three-word phrase output. Those phrases selected by a subject analyst for inclusion in the KB are indicated with asterisks. The second column lists the output that the KBB program identifies as being single words. Several acronyms and material abbreviations have been recognized and flagged by a subject analyst.

Single-Word Term Assessment Tool

One early problem with NASA's MAI system was the preponderance of single-word thesaurus terms that the system generated. About 40% of NASA's Thesaurus terms are single-word terms; however, indexers in the NASA environment tend to use single-word terms only about 20% of the time. In an aerospace database, the term AIRCRAFT, for example, is too general for helpful indexing. An assessment tool was designed to improve the computer-generated set of index terms by reducing the number of single-word terms inappropriately suggested by MAI. See Appendix D for the specifications. The procedure identified the single-word thesaurus terms that frequently occurred in text but were seldom used by human indexers. The terms were sorted and listed by percentages that indicated how often indexers assign one-word terms that appeared in titles and abstracts. Figure 16 shows a sample of an actual printout used by the analysts to reduce the number of single-word thesaurus terms suggested by MAI. Those terms with simultaneously high text counts and low percentages were candidate terms for null translations. In some cases, translations are qualified by the addition of a second word.

Un-edited KBB Output for METAL MATRIX COMPOSITES

THREE-WORD PHRASE OUTPUT	SINGLE WORD OUTPUT
482 * METAL MATRIX COMPOSITE(S)	74 FIBER-MATRIX
72 BEHAVIOR OF COMPOSITES	70 * MMCS
72 STRENGTH OF COMPOSITE(S)	47 REINFORCEMENTS
62 * REINFORCED METAL MATRIX	45 FIBER / MATRIX
61 * ALUMINUM MATRIX COMPOSITE(S)	41 * SIC / AL
55 PROPERTIES OF COMPOSITES	29 STRENGTHENING
51 REINFORCED MATRIX COMPOSITE(S)	29 UNREINFORCED
49 * REINFORCED METAL COMPOSITE(S)	27 * BORON / ALUMINUM
48 * REINFORCED ALUMINUM COMPOSITE(S)	25 MODULI
46 FIBER AND MATRIX	24 * GRAPHITE / ALUMINUM
42 * FIBER REINFORCED METAL(S)	21 * AL-SIC
40 FIBER MATRIX COMPOSITE(S)	20 FP
38 BEHAVIOR OF MATRIX	19 STRENGTHENED
37 BEHAVIOR OF METAL	18 MICROGRAPHS
35 FIBER REINFORCED MATRIX	17 * AL-MATRIX
33 * FIBER METAL COMPOSITE(S)	17 * ARALL
33 * FIBER REINFORCED ALUMINUM	16 ADDITIONS
33 PROPERTIES OF REINFORCED	16 EXTRUDED
32 PROPERTIES OF MATRIX	16 FRACTOGRAPHIC
31 * FIBER METAL MATRIX	16 * GR / AL
30 PROPERTIES OF METAL	16 * GR / MG
29 * ALUMINUM ALLOY MATRIX	16 PARTICULATE-REINFORCED
29 FIBER VOLUME FRACTION	16 SIC-REINFORCED
29 STRENGTH OF FIBER(S)	15 * AL-SI
28 * ALLOY MATRIX COMPOSITE(S)	14 * AL / SIC
28 * ALUMINUM ALLOY COMPOSITE(S)	
28 CHARACTERISTICS OF COMPOSITE(S)	
28 PROPERTIES OF ALUMINUM	
28 PROPERTIES OF FIBER(S)	
27 * METAL MATRIX MATERIAL(S)	
26 * SILICON CARBIDE ALUMINUM	
24 PROPERTIES OF ALLOY(S)	
24 * SIC REINFORCED ALUMINUM	
23 * ALUMINUM METAL MATRIX	
22 BEHAVIOR OF ALUMINUM	
22 HIGH TEMPERATURE COMPOSITES	
22 * REINFORCED ALUMINUM ALLOY(S)	
22 SILICON CARBIDE WHISKER(S)	
22 TRANSMISSION ELECTRON MICROSCOPY	
21 * FIBER ALUMINUM COMPOSITE(S)	
21 THERMAL EXPANSION COEFFICIENT(S)	
20 * CARBIDE REINFORCED ALUMINUM	
20 FATIGUE CRACK GROWTH	

Figure 15

ONE-WORD THESAURUS TERMS THAT ARE CONTAINED IN TEXT

THES	TEXT	TEXTCNT	STICNT	PERCT
ABORIGINES	ABORIGINE	2	0	0.00
ABORIGINES	ABORIGINES	1	0	0.00
ACCLIMATIZATION	ACCLIMATIZATIONS	1	0	0.00
ACETONE	ACETONES	1	0	0.00
ACETONITRILE	ACETONITRILES	1	0	0.00
ACETYLACETONE	ACETYLACETONES	1	0	0.00
ACRIFLAVINE	ACRIFLAVINE	2	0	0.00
ACRYLONITRILES	ACRYLONITRILES	1	0	0.00
AEROBES	AEROBE	1	0	0.00
ALKYLFERROCENE	ALKYLFERROCENE	1	0	0.00
~~ALLOWANCES~~	ALLOWANCE → oo	5687	7	0.00
ALLOWANCES	ALLOWANCES	172	0	0.00
ALTERNATIONS	ALTERNATIONS	43	0	0.00
ALUM	ALUMS	2	0	0.00
AMBIENCE	AMBIENCE	18	0	0.00
AMOEBA	AMOEBAS	1	0	0.00
ANASTIGMATISM	ANASTIGMATISM	1	0	0.00
ANGIOSPERMS	ANGIOSPERMS	6	0	0.00
ANTIADRENERGICS	ANTIADRENERGIC	1	0	0.00
ANTICHOLINERGICS	ANTICHOLINERGICS	1	0	0.00
ANTIHISTAMINICS	ANTIHISTAMINIC	1	0	0.00
APHELIONS	APHELIONS	2	0	0.00
APPEARANCE	APPEARANCES	130	0	0.00
ARCHIPELAGOES	ARCHIPELAGOES	1	0	0.00
~~ARTS~~	ART → oo	8163	10	0.00
ASSUMPTIONS	ASSUMPTION → oo	8936	9	0.00
ASSUMPTIONS	ASSUMPTIONS	7939	14	0.00
ASTROGRAPHY	ASTROGRAPHY	1	0	0.00
ATTRACTION	ATTRACTIONS	73	0	0.00
ABSTRACTS	ABSTRACT → oo	23659	185	0.01
ACCOMMODATION	ACCOMMODATIONS	220	3	0.01
ACHIEVEMENT	ACHIEVEMENT → oo	1326	15	0.01
ACTIVATION (BIOLOGY)	ACTIVATION → oo	5879	41	0.01
AMPLIFICATION	AMPLIFICATIONS → oo	81	1	0.01
ACUITY	ACUITY	470	11	0.02
ANTIMONIDES	ANTIMONIDE	371	8	0.02
ARSENIDES	ARSENIDE	2240	42	0.02
ATTACHMENT	ATTACHMENTS → oo	340	6	0.02
ACTIVATION	ACTIVATION → oo	5879	171	0.03
ADJUSTING	ADJUSTING	1343	43	0.03
ALTERNATIVES	ALTERNATIVES	3588	97	0.03
ALTIMETRY	ALTIMETRY	752	19	0.03
ALTITUDE	ALTITUDES	5919	177	0.03
ARTS	ARTS	194	5	0.03
ATTITUDE (INCLINATION)	ATTITUDES	723	21	0.03
ATTRACTION	ATTRACTION	750	26	0.03
ACCUMULATORS (COMPUTERS)	ACCUMULATORS	188	8	0.04
ALLOYS	ALLOY	22555	847	0.04
ATOMS	ATOM	5024	205	0.04
ATOMS	ATOMS	10261	389	0.04
ATTITUDE (INCLINATION)	ATTITUDE	7913	305	0.04
AVERAGE	AVERAGES	1660	72	0.04
ACCESSORIES	ACCESSORIES	247	13	0.05
ACCIDENTS	ACCIDENTS → oo	2081	103	0.05

fine; arts; people → oo
" ; " ; oo → oo
999

Indexer Feedback

Another important aspect of building and improving the KB is indexer feedback. Indexers are urged to provide both missing translations, whenever they find any, and corrections to existing translations.

On some occasions, translations that are erroneous for the document at hand are not changed. The reason for this is that KB construction has been based, in part, on probabilities. If the given input can be translated in more than one way, the analyst considers the following:

- What is in the NASA STI database? For example, if the word in question is "blowouts" and in the NASA database this word always refers to airplane tires, there is no need to be concerned with an alternate meaning related to oil or gas wells.

- If the term has more than one meaning evident in the database, can it be qualified with the addition of another word or words? For example, do the phrases "tire blowouts" or "well blowouts" both occur? Are there other equivalent expressions that occur? If so, several or many additions may need to be made to the KB to clarify when one translation will be provided and when a different translation is more appropriate.

- Is the ambiguous input ambiguous because the meaning has changed over a period of time? If so, most documents that will be indexed in the future will require the newer meaning.

- If there is still a question as to how to translate the input, and the concept is important for the meaning that occurs most of the time (say 80%), that meaning will be entered as the term for MAI to suggest. Indexers would need to delete that term only when it is inappropriate (or 20% of the time).

The analyst must decide what the cutoff probability will be for each entry in question. For a word or phrase that occurs frequently the percentage may be higher than for a word or phrase that has only a few occurrences. Also the indexing importance of the concept should weight the decision. The probability that the indexer would assign a term if suggested by MAI is expressed as the ratio of the number of times the indexer used the term when the concept appeared in text to the number of times the word or phrase occurred in the text.

DATA FILE MAINTENANCE

Two sources of information must be processed regularly to update the MAI KB. These sources are:

- Updates to the NASA Thesaurus.
- Modifications and additions suggested by NASA indexers.

Updates to DTIC and DOE thesauri must also be examined. As long as Subject Switching is being used, the Subject Switching files must be maintained. Updates to the DTIC-to-NASA Subject Switching file are described in NASA-CR-3838. DOE-to-NASA updating is handled in the same way, substituting the DOE prefix for DTIC (Silvester, Newton, and Klingbiel, 1984). Natural language should also be researched on RECON for DTIC and DOE subject terms, and when appropriate, these should be added to the MAI KB.

Online maintenance commands are described at length in Appendix E.

RESULTS AND CONCLUSIONS

Applications

The NASA Lexical Dictionary system can be used for any application that requires the identification of NASA Thesaurus terms from equivalent natural language words and phrases, or vice versa. Its primary purpose has been to generate NASA Thesaurus terms from any designated input text - usually titles and abstracts - and to present these terms to the indexer for acceptance or rejection. This, however, is not the only use that has been found for the system.

Another application for MAI is its use as a custom-built spell-checker. The KB has been expanded intentionally beyond the needs of MAI in order to improve this spelling check feature, and further additions are planned. Words that cannot be found as the first element of a key often cannot be found because they are misspelled.

Machine aided indexing has also been used to generate candidate NASA Thesaurus terms from the Library of Congress' subject headings. Another project used MAI to re-index more than 300,000 records in the NASA database that were indexed before the controlled vocabulary existed. The re-indexing results have been reviewed and the machine selections are generally good. A separate report, entitled "Automatic Re-indexing of NASA's Pre-thesaurus STI Records," is being prepared to document this experience, and will be submitted for publication as a separate NASA Contractor Report.

MAI is used by the Thesaurus Lexicographer to identify Thesaurus terms in term definitions. Any identified terms that are defined are printed in bold type when they appear in definitions, thus providing a cross reference capability not possible without MAI.

Potential applications, not yet explored, are as a front end for searching files in the RECON system, as an indexing aid for machine readable full text, and generating NASA Thesaurus terms for the records in the National Advisory Committee for Aeronautics (NACA) collection.

Impact Evaluation

In NASA's high-pressure production environment, testing the MAI system, without slowing the work that must be done, has presented a challenge. The number of indexers currently abstracting and indexing has been reduced from 8 to 5 since the institution of MAI; however, MAI is not the only change responsible for increased productivity. Input is now done at a computer terminal or electronically from magnetic tape instead of with pen and paper. This also speeds processing. The fact remains that the size of the present indexing staff is about 62% of the pre-MAI staff size and the individual's output has approximately doubled in the past 10 years.

Other variables that can affect the measures of the system include:

- The amount of time available to an indexer. MAI terms may be questioned less if the workload is heavy and more if the load is light.

- The existence of similar terms, for example SIMULATORS and SIMULATION. The indexer may select one when MAI suggests the other. Index terms that express such closely related concepts could be counted as the same to indicate that MAI has suggested an appropriate term.

- Valid terms that were suggested by MAI, appropriate for the document at hand, and not assigned by the indexer.

- Terms that were assigned by the indexer and that should not have been.

It was determined in an early test that machine aided indexing saved an average of 3 minutes per document by reducing the time needed to look up terms in the thesaurus (Silvester, Newton, and Klingbiel, 1984). It is reasonable to expect that this time savings is even greater for comparatively new indexers, which refers now to about 40% of the staff.

Match Rate

The first measure of how well the MAI system performed was referred to as the match rate. In the migration of this measure from DTIC to NASA, its meaning changed. At DTIC, Klingbiel used it to describe the percentage of machine selected phrases that were either partially or completely matched. At NASA the match rate referred to the percentage of MAI suggested terms that the indexer elected to use. By the new definition, this measure has ranged from 23% to 75%, depending upon the file being processed and other variables. Typically, the rate tends to rise gradually as improvements are made to the system, and then to drop as indexers become used to it and become more demanding and selective in its use.

Capture Rate

In November of 1986, NASA instituted another measure referred to as the capture rate. This described the percentage of indexer-assigned terms that were suggested by MAI. The capture rate has been, rather consistently, a few percentage points higher than the match rate.

Consistency Factor

A third measure, which we began to use in September 1989, was a consistency (or quality) factor "q". This measures the percentage of common terms "c" found in two lists of terms, one generated automatically and representedd by "a", and the other terms selected intellectually by the indexer and represented by "i". Expressed in another way, "q" is the ratio of the common terms to the unique terms, where $q = c/(a+i)-c$ (Lustig & Knorz, 1986; Lancaster, 1991).

In a recent test of one of NASA's files, which in full text consists of only abstracts and titles, the above measures were as follows:

Year	Match Rate	Capture Rate	Consistency Factor	
1983	26.1 %	45.7 %	19.9 %	
1986	23.4	42.9	17.8	
1987	62.8	77.6	53.2	(Sample 1; Access-1)
1987	65.6	69.2	50.8	(Sample 2; Access-2)
1991	79.1	80.1	65.3	

These figures are for 5 different samples of 30 documents each, the minimum sample size for valid results for the population.

Benefits

Several benefits come from the use of MAI.

- o MAI-suggested terms do not have to be verified as to the correctness of form and spelling.
- o Unwanted terms can be eliminated with a single stroke as opposed to the typing (or writing) of up to 42 characters to add a term.
- o Appropriate, unfamiliar, technical terms may be suggested which the indexer would omit without a prompt from MAI.
- o Indexer research time is reduced because MAI entries are researched before their addition to the KB, thereby presenting expert advice.

- The increased number of indexing terms that often results from using MAI provides additional access points for records.

- MAI-suggested terms function as a check-list of indexable concepts and increases the consistency of indexing.

- Spinoffs from MAI provide aids for the Proofreaders, the Thesaurus Lexicographer, and the Retrieval Analysts.

PROJECT SUMMARY

In the NLD, the NASA STI Program has a system designed to translate natural language words and phrases from any source into equivalent concepts expressed in NASA posting terms. However, the Knowledge Base can be used for any application that requires the identification of equivalent NASA terms and natural language words and phrases. The system was initiated to allow reuse of DTIC indexing in the NASA environment. It was expanded to include the indexing of DOE as input, and finally, to accept any designated text or terms. As the size of the Knowledge Base increases with carefully and statistically selected entries, the output becomes more acceptable to the indexers. However, as the output improves, the indexers become more demanding of the system.

The current implementation of natural language MAI at CASI begins with the input of document records, particularly the titles and abstracts. Some records are received in machine-readable form on magnetic tape. At least four of these are run each month on appropriate programs in a batch mode. Other records are typed by CASI staff into the online Input Processing System (IPS). IPS is mounted on an IBM 4381 mainframe. It is accessed by input processing staff from an IBM 3278-4 or 3180-type terminal. NASA abstractor/indexers (A/I) who type in abstracts have two options for accessing natural language MAI. They can use MAI online, in an interactive mode, or they can transfer the document records to a queue, which is processed through MAI in a batch. This batch processing is used if input is supplied by non-A/I support personnel. MAI batches are run four times a day. An indexer who wants to use interactive MAI presses a function key, and, within approximately 7 seconds for a title plus a 150-250 word abstract, receives 10-15 thesaurus terms for consideration. The indexer reviews these candidate terms and makes additions or deletions as needed. The system currently serves five indexers but has been stress-tested with several more.

The MAI KB today has more than 113,000 records, occupying 581 tracks, or nearly 28 megabytes of storage. Although the KB growth rate is now declining, the file could grow to around 180,000 or even 200,000 entries with the analysis of text targeted to additional, highly-posted thesaurus terms.

In addition to using the NLD system for an indexer's aid, several other uses have evolved. Proofreaders, the Thesaurus Lexicographer, and the Retrieval staff have tools that are spinoffs of the NLD system, and new uses and enhancements are thought up faster than they can be tried. Several goals remain for the MAI project. One is to provide the indexers who use MAI online with MAI-suggested terms in less than 1 second. Another is to rank the suggested terms in order to speed up the designation of major and minor terms. When these challenges have been met, there will most likely be new ones to keep NASA's MAI system at the forefront of natural language processing.

GLOSSARY

Access-1 - NASA's original, general-purpose, computer program that accesses the NLD MAI and Subject Switching files. This program never operates independently; it is always called by an application program.

Access-2 - NASA's revised, general- purpose, computer program that delineates phrases and accesses the KB. This program, like Access-1, never operates independently, but is always called by an application program.

DTIC - Defense Technical Information Center.

IPS - NASA's electronic Input Processing System.

KB - Knowledge Base.

KB key - a unique word or phrase of input for which an equivalent NASA Thesaurus term is (or terms are) sought; used as a Knowledge Base record's address in the computer.

Knowledge Base - the master file (in matrix form). It accepts input in natural language and provides output in NASA Thesaurus terms that express the same concept.

MAI - machine aided indexing.

MAPS - MAchine Phrase Selection. This is a computer program that consists of several subroutines and functions whose main purpose is to analyze the input text and break it down into phrases.

NASA Lexical Dictionary - (1) A system for generating NASA Thesaurus terms automatically from any specified input. (2) The Knowledge Base. See also page 5, "Definition."

natural language - English as it is written.

NLD - NASA Lexical Dictionary.

phrase delineation - a procedure for breaking natural language text into strings of words usually shorter than a sentence.

search key - two or more words that occur within a single natural language phrase, that are concatenated to form a possible KB key, and that, if found, point to the same concept expressed in NASA Thesaurus terms. See also "semantic unit."

semantic unit - a group of natural language words that express an indexable concept. See also "Search key."

stopwords - words without indexable content that occur frequently and are used to break text into strings of words shorter than a sentence.

string - a group of sequential words in natural language text.

APPENDIX A: THE NASA RECOGNITION DICTIONARY (NRD)

PURPOSE

The purpose of the NRD was to assign syntactic class (category code) numbers to words encountered in text to be processed by the Machine Phrase Selection (MAPS) program. See Appendix B for the grammar rules.

CATEGORIES

The categories, which represent syntactic classes, completely determined the roles of the words in the phrase selection process. MAPS used these category numbers to identify and print out "acceptable" phrases.

The 14 categories used by NASA's MAPS program are summarized in Table A-1 and are listed below in 4 sets according to their functions and the degrees to which they were involved in the routine maintenance of the NRD. It will be noted that the category numbers are not entirely consecutive. This is because the NASA MAI system was patterned after DTIC's system. The "missing" category numbers were used by DTIC, but did not appear to be needed by NASA. They were reserved in case they were required at a later date.

Set 1

Category 01 - stopwords, marked by the letter S. Words in this category did not appear in the MAPS output. Originally, these were words that DTIC had decided were without indexable content. They served to break text into strings called phrases (hopefully noun phrases). In mid 1985, about 55% of the words were in this category and the number was growing. Since the object of the system was to identify noun phrases with indexable concept, this category consisted of a variety of verbs, adverbs, some proper nouns, pronouns, ambiguous acronyms, gerunds, and anything that was considered to be without indexing value.

Set 2

This was the most frequently used set of categories, and new words not marked category 01 most often fell into one of the 3 categories of this set.

Category 02 - an adjectival category, marked by the letter A. Nearly 14% of the words in the dictionary belonged to this group.

Category 03 - a strong noun category, marked by the letter N. Category 03 was the ONLY group of words whose members could occur in isolation. As such, it was the ONLY single word output permitted by MAPS. About 18% of the words in the dictionary belonged to this category.

Category 06 - a weak noun category, marked by the letter Z. Words in this group could NOT stand alone. They could combine with each other, or with members of the 02 or 03 categories. About 11% of the words in the NRD belonged to this group.

Set 3

Words were assigned to a category in this set when they did not fall into Set 1, Set 2, or Set 4.

Category 14 - a very restricted category of adjectives, marked by the letter B. These words could not appear in indexable isolation or occur in final position, and were required to combine with words in category 21, i. e., "A" and "BE". Only a few words belonged to category 14.

Category 16 - a positionally restricted category of weak nouns, marked by the letter H. This class was created to distinguish between verbal and nominal use of the traditional gerund. The words in this class, when occurring either medially of finally, tended to be nouns. About 1% of the words (1,250) fell into this category.

Set 4

The following categories were established for very special circumstances and, for all practical purposes, were closed.

Category 8 - marked by the letter P. This category was designed to handle the following prepositions and conjunctions: OF as in the heat of formation, TO as in air to air, AND, BETWEEN, FOR, FROM, IN, ON, OR, and PER.

Category 11 - comma. This category was needed to handle a series of adjectives or nouns.

Category 17 - marked by the letter C. This category consisted of the single word OVER.

Category 18 - marked by the letter T. This category consisted of the single word THE.

Category 19 - marked by the letter D. This category consisted of 4 stand-alone nouns that are context sensitive with regard to the words A or BE. These words are Canada, HEAO, ISEE, and stars.

Category 20 - a restricted group of context sensitive weak nouns, marked by the letter F. This category consisted of the following words: Agena, Bomarc, HEOS, Magsat, and Palapa.

Category 21 - marked by the letter G. This category consisted of two words: A and BE. Category 21 was intended to be used before a D (Category 19), after a B (Category 14), or after a J (Category 22) word, very restricted sets of nouns and adjectives.

Category 22 - a restricted group of context sensitive weak nouns, marked by the letter J. The category consisted of the following words: Explorer, observatory, orbiter, satellite, and stage.

Table A-1

Word Category	Letter Designation	Set No.	Category Description
01	S	1	Stopwords
02	A	2	Adjectival words
03	N	2	Strong nouns
06	Z	2	Weak nouns
08	P	4	Prepositions and conjunctions
11	,	4	Comma
14	B	3	Restricted, adjectival words
16	H	3	Restricted, weak nouns
17	C	4	Single word: over
18	T	4	Single word: the
19	D	4	Restricted to: Canada, HEAO, ISEE, and stars. Strong nouns; context sensive in regard to "A" and "BE".
20	F	4	Restricted to Agena, Bomarc, HEOS, Magsat, and Palapa. Context sensitive weak nouns.
21	G	4	Two words: A and BE.
22	J	4	Restricted to: Explorer,observatory, orbiter, satellite, and stage. Context sensitive weak nouns.

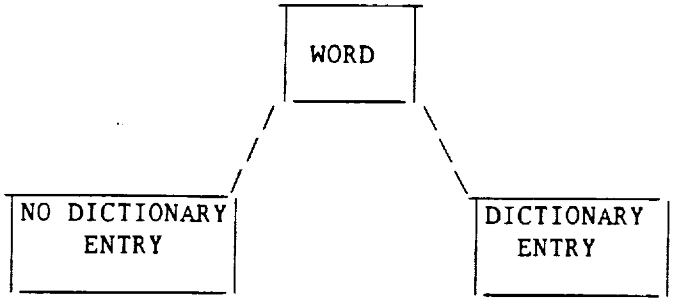
Determining the Category of a Word

In order to determine a word's category, it was important to see how that word occurs in text in the NASA environment. This was done by searching the NASA database for documents that had the word in question in the title. The rule of thumb required a minimum of 8 documents with the word in the title, or a KWIC index, or information from some other database as a body of knowledge upon which to base the category code assignment.

When the desired word was found in context, a series of binary decisions were made, These will be illustrated as a decision tree. The first decision was whether or not the word was to be added to the NRD. The following classes of words were not added:

Misspellings
Foreign, nontechnical terms
Pure Arabic numerics
Alphanumerics
Personal nouns unless integrally associated with equipment, processes, efforts, etc.
Unique or idiosyncratic abbreviations and acronyms
Words containing hyphens or other symbols

All other words were added to the NRD. The process to this point was:

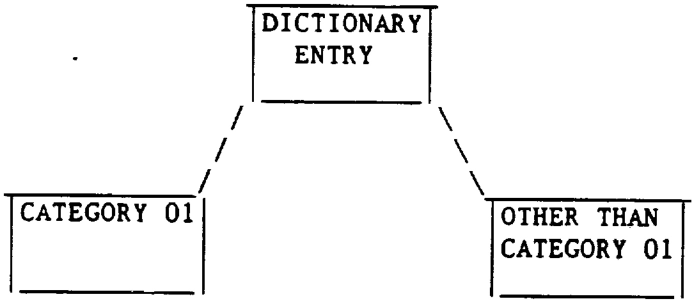


Category 01

Dictionary entries were of two types: those that contributed to the formation of indexing phrases and those that did not. The words that did not contribute were coded as Category 01. At DTIC, Klingbiel examined and coded a body of 4 million words of text (about 40,000 unique words) and found that approximately half were designated as Category 01. The crucial decision in Category 01 assignments is in the significance of the candidate word for the indexing and retrieval process. Only nonsignificant words should be marked Category 01. In general NASA did not utilize for indexing and retrieval those words that are (almost always) verbs, adverbs, interjections, or pronouns. Most function words are also Category 01: BUT, NOT, AN, EITHER, etc.; however, we did retain as useful for indexing: AND, BETWEEN, FOR, FROM, IN, OF, ON, OR, PER TO, and THE under very restrictive conditions.

Stated another way, the primary traditional word categories useful for indexing are nouns and adjectives. These words became Category 01 only when they bore no essential technical content either alone or in combination. Such words as "easier," "directive," "crucial," and "explanation" are in this category.

Finally, if the word were not to be coded Category 01, then it was a word that contributed to the indexing and retrieval process, and it was coded as one of the active categories. The decision tree branches as follows:



When it was decided that a word was to be entered into the dictionary, and that the designation was other that Category 01, there were 5 other categories to choose from: two adjectival and three noun types. The chief distinction between these two groups was positional. Adjectives are never allowed to occur in isolation, nor are they allowed to occur in the final position.

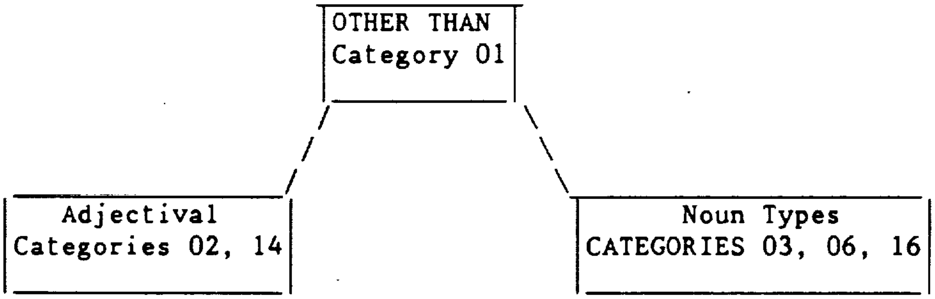
To begin the process of choosing a category other than 01, we studied the occurrences of the candidate word.

- If the word occurred in indexable isolation in the sense that its immediate predecessor and immediate successor were Category 01 words, we assigned Category 03, provided the word was a single-word NASA Thesaurus term, or translatable into one or more NASA Thesaurus terms, or a strong candidate for a new NASA Thesaurus term.

- If the word accurred in word final position of two word (or longer) phrases with desirable technical content, the word could not be an adjective; therefore the category COULD NOT be either 02 or 14.

- If the word were positionally sensitive (occurred only initially and/or medially, for instance, or occasionally in isolation), the word was a candidate for an adjectival designation. If the occasional occurrence in indexable isolation was such that the word should be captured, the word MUST be designated Category 03, otherwise the adjectival designation was still a possibility.

The branching process continues as follows:



Adjectival Categories

Category 02

This category contains a very large group of words that, for MAPS, acted as adjectives. All of the words in this class are positionally sensitive. No "adjective" could occur in final position in an MAI phrase, nor could any word marked as an adjective appear in isolation. Consequently, an adjective (Category 02) is ALWAYS part of a phrase. Since this category is based on position, it is unnecessary for an 02 word to have an -al, -ic, -ed, or other standard adjectival suffix.

Category 14

The words in this group are also adjectives and therefore may not appear in indexable isolation, nor in the final position. They differ from Category 02 adjectives in that they are desired only in very limited contexts.

The 5 words in this class were identified as words that combine with the letter "A", Category 21. "BE" was later added to Category 21 to take care of documents on BE stars. Adding words to Category 14 that did not combine with "A" or "BE," would have required the setting up of their desired co-occurrents in Category 21. The Category 14 words were: ANIK, CASSIOPEIA, COMPOUND, HELIOS, and VITAMIN.

The Noun Classes

When a word to be added to the dictionary did not fit the adjectival requirements, the word was a candidate for one of the noun categories:

- o Category 16 consisted of words that were not allowed in initial position.
- o Category 03 consisted of words that were strong nouns, that is, words that were wanted even when they occur in indexable isolation.
- o Category 06 was used if neither of the above categories applied.

Category 16

This category marked "weak nouns" that were positionally sensitive and were restricted to the medial or final position in a phrase. Category 16 was created in an attempt to distinguish certain verb-noun ambiguities, one of which is the gerund form identified by the suffix "-ing." As an empirical fact, it appears that many of these verb-noun ambiguities are disambiguated when the word in question is limited to the medial and/or final phrase position. That is, observation shows that in many cases a given word form is a verb in initial position, but a noun form in medial or final position. If it is observed that a word in initial position functions as a verb, it should be assigned to this group. Word assignments to Category 16 were not based on an "-ing" ending, but on the restriction of the word to the medial or final position in a phrase.

Category 03

This category marked strong nouns. Words in this category are NOT positionally sensitive. This is the only class of words whose members may stand in indexable isolation, that is, they may stand alone and express an indexable concept. Words were not assigned to this class unless they were semantically strong enough to be sensible retrieval terms in isolation. A word that is not of interest in every one of its instances should not be assigned to this category.

Category 06

This category marked weak nouns that are not positionally sensitive. These nouns are weak in a semantic sense, in that they are not useful in isolation. They are useful only when modified or if they modify some other word. All nouns that did not require the special attributes for Category 16 or 03, belonged to Category 06. Also all hyphenated words were assigned the 06 category.

Category Policy

The category policy may be summarized as follows:

1. Do not assign category numbers to any:

   o Misspelled terms
   o Combined words that should be separated
   o Unknown or unidentified words (in terms of meaning)
   o Foreign, nontechnical words
   o Alphanumerics
   o Hyphenated words or words that contain a virgule, ampersand, or other embedded symbol

2. Assign Category 01 to nonsignificant words including:

   o Ad hoc abbreviations and acronyms
   o Incidental names of cities, towns, geographical locations
   o Incidental proper names

3. Assign Category 02 to words that function like adjectives including:

   o Chemical terms ending in -yl, -ic, -o, or -i
   o Chemical radicals
   o Personal names associated with mathematical equations, physical, chemical, or biological laws, physical, chemical, or biological processes

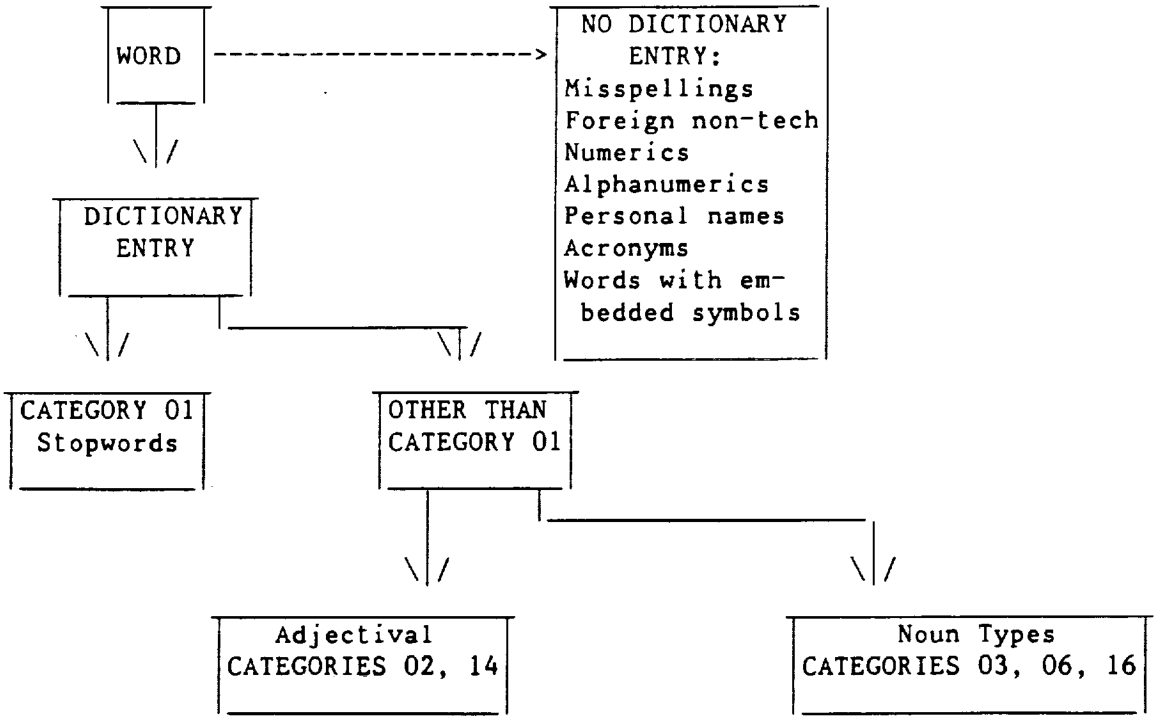
4. Assign Category 06 to words that are not positionally sensitive, and may not stand alone including:

   o Identifier names like -operations, -project, -satellite, etc.
   o Biological species
   o Singular chemical terms not useful in isolation

5. Assign Category 03 only to semantically strong, unambiguous nouns including:

   o Biological phylum, class, order, family, genus
   o Plural chemical terms useful in isolation

6. Assign categories using the following complete decision tree:



ONLINE MAINTENANCE SYSTEM COMMANDS

Command Overview

The NRD system provides a series of commands for use in maintaining the NRD file. Command names begin with NRD and are similar in names and functions to some of the commands used for Knowledge Base maintenance. The NRD commands are as follows:

Command Name:	Function:
NRDFIND	Displays NRD entries online
NRDLOAD	Loads update transactions into the master NRD file
NRDPRNTC	Prints the NRD, sorted first by category, then by word
NRDPRNTN	Prints the NRD, sorted by word and showing the number of times the word has been encountered in text by MAPS
NRDPRNTW	Prints the NRD, sorted by word
NRDUPDT	Allows the entry of update transactions

Command Descriptions

NRDFIND. This command searches the NRD for a specific word and prints at the terminal ten sequential NRD records, beginning with the word requested, if it exists. If the requested word is not found, the program locates the sequential position in which the word should occur and prints the next ten records.

NRDLOAD. This command loads additions and corrections from a working dataset into the master file. When additions or corrections are needed in any NRD record, the update (NRDUPDT) command is used to create this dataset. Dataset editing is done online, using the SPF (System Productivity Facility) editing capability, before the transactions are loaded into the master file.

NRDPRNTC; NRDPRNTW. These commands generate prints of the master file: C is sorted by category (see Figure A-1) and W sorted alphabetically by word (see Figure A-2). A sort by number of occurrences (count), PRINTN, was requested and programmed but never debugged; however, the count appears in both of the other sorts.

NRDUPDT. This command puts the changes, additions, or deletions into a dataset that is editable online. It also provides some checks for format. NRD records are entered into the work dataset in the following format:

Column 1 - flag (must be blank or #)

- The flag is a pound sign # if the word occurred in any NASA Thesaurus term or Use reference.
- The flag is a blank for any word not found in the NASA controlled vocabulary including Use references.

Column 2-3 - category of the word expressed by two numeric digits or, to delete a record, two asterisks (**).

Column 4-? - word (no embedded blanks or special characters); all alphas.

In making changes to the category of a word, it may be necessary to manually step through the programming for phrase selection. See Appendix B.

NRD commands are executed on the MAI Project Coordinator's IBM 3180 terminal which is connected to two IBM 4381 mainframes. Using TSO (Time Sharing Option), commands are entered following the computer's prompt of READY. Archived datasets must be restored before these commands can be executed.

```
                         NASA RECOGNITION DICTIONARY IN SEQUENCE BY CATEGORY

*02 WIREBASED                0  #02 WRAPAROUND                 0  *02 YTTRIATED                 0
*02 WIREBOND                 0  *02 WRAPPABLE                  0  *02 YUGOSLAVIAN               0
*02 WIREBONDABLE             0  *02 WRAPPED                    0  #02 YUKAWA                    0
*02 WIREBONDED               0  *02 WRECKED                    0  #02 YUKON                     0
*02 WIRED                    0  *02 WRINKLED                   0  #02 YURTUK                    0
*02 WIREFREE                 0  #02 WROUGHT                    0  #02 ZEEMAN                    0
*02 WIREGUIDED               0  #02 X                         26  #02 ZENER                     0
#02 WIRELESS                 0  #02 XANTHIC                    0  *02 ZENKERS                   0
*02 WIREWAY                  0  *02 XENOBIOTIC                 0  *02 ZENON                     0
*02 WIREWOUND                0  *02 XENOGENEIC                 0  #02 ZERO                     20
*02 WIREWRAPPED              0  *02 XENOGENIC                  0  #02 ZETA                      0
*02 WIRTINGER                0  *02 XENOLITHIC                 0  #02 ZEUS                      0
*02 WISHFUL                  0  *02 XERO                       0  #02 ZIEGLER                   0
*02 WISNA                    0  *02 XEROGRAPHIC                0  *02 ZIP                       0
#02 WISWESSER                0   02 XEROPHYTIC                 0  #02 ZIRCALOY                  0
*02 WITCH                    0   02 XEROPHYTICALLY             0  *02 ZIRCONYL                  0
#02 WKB                      0  *02 XEROXING                   0  #02 ZODIACAL                  0
#02 WOODEN                   0  #02 XI                         0  #02 ZONAL                     0
*02 WOODY                    0  *02 XYLOASCORBIC               0  #02 ZOND                      0
*02 WOOLEN                   0  #02 Y                          3  *02 ZOOGEOGRAPHICAL           0
#02 WORD                     0  *02 YABA                       0  *02 ZOOLOGICAL                0
*02 WORKER                   0  *02 YACHTMANS                  0  *02 ZOOM                      0
#02 WORKHORSE                0  #02 YAGI                       0  *02 ZOONOTIC                  0
*02 WORLDWIDE                7  #02 YAK                        0  *02 ZOOPLANKTONIC             0
*02 WORMLIKE                 0  *02 YARD                       0  *02 ZOSCHKE                   0
*02 WORN                     0  *02 YAWED                      0  #02 ZPR                       0
*02 WORSTED                  0  #02 YAWING                     0  #02 ZUNI                      0
#02 WOUND                    0  *02 YELLOWISH                  0  *02 ZWITTERIONIC              0
*02 WOUNDING                 0  #02 YELLOWSTONE                0  *02 ZYGMUND                   0
*02 WOVEN                   11  *02 YOKED
#02 WRANGELL                 0  #02 YOUNG


                         NASA RECOGNITION DICTIONARY IN SEQUENCE BY CATEGORY

*03 ABACA                    0  *03 ACARICIDES                 0  *03 ACETOPHENONES             0
*03 ABATTOIRS                0  *03 ACCELERATIONS              0  *03 ACETOPHTHALATE            0
#03 ABDOMEN                  0  *03 ACCELEROGRAMS              0  *03 ACETOXIME                 0
*03 ABDOMENS                 0  *03 ACCELEROMETER              0  *03 ACETOXYBENZENE            0
#03 ABERRATION               0  #03 ACCELEROMETERS             4  *03 ACETOXYPIPERIDINIUM       0
*03 ABERRATORS               0  *03 ACCENTUATION               0  *03 ACETOXYSILANE             0
*03 ABIETATE                 0  #03 ACCEPTABILITY              0  *03 ACETOXYSILANES            0
#03 ABILITIES                3  #03 ACCEPTANCE                12  *03 ACETYLACETONATE           0
#03 ABIOGENESIS              0  *03 ACCEPTORS                  0  *03 ACETYLACETONATES          0
#03 ABLATION                 0  *03 ACCESSIBILITY              0  #03 ACETYLACETONE             0
*03 ABLATIVES                0  #03 ACCESSORIES                0  *03 ACETYLACETONES            0
*03 ABLATIVITY               0  #03 ACCIDENTS                 22  *03 ACETYLASE                 0
*03 ABLATORS                 0  *03 ACCLIMATION                0  #03 ACETYLATION               0
#03 ABLESTAR                 0  #03 ACCLIMATIZATION            0  *03 ACETYLCARNITINE           0
#03 ABM                      0  #03 ACCOMMODATION              4  *03 ACETYLCELLULOSE           0
*03 ABMS                     0  *03 ACCOUNTABILITY             0  *03 ACETYLCHOLINE             0
#03 ABNORMALITIES            0  *03 ACCOUNTANT                 0  *03 ACETYLCHOLINESTERASE      0
```

Figure A-1

NASA RECOGNITION DICTIONARY IN SEQUENCE BY WORD

```
*01 LALP                   0  *03 LAMINECTOMY          0  *06 LANDERS             90
*01 LALR                   0  *06 LAMINIA              0  *01 LANDFAE              0
*01 LALS                   0  *02 LAMINOGRAPHIC        0  *06 LANDFALL             6
*01 LALSD                  0  *03 LAMINOGRAPHY         0  *06 LANDFILL             4
*01 LAM                    9  *01 LAMMERS              0  *01 LANDFILLED           0
*01 LAMA                   0  *01 LAMMR                0  *06 LANDFILLING          0
*01 LAMALLOY               0  *01 LAMONT               0  #03 LANDFILLS            2
*01 LAMAR                  4  *01 LAMOR                2  *06 LANDFORM            47
*01 LAMARCK                0  *01 LAMOUR               0  #03 LANDFORMS          246
*01 LAMARS                 1  *01 LAMOURE              0  *01 LANDFRIED            0
*01 LAMAS                  0  *06 LAMP               233  *01 LANDI                0
*01 LAMAX                  0  *01 LAMPAC               0  #03 LANDING           1246
#02 LAMB                  38  *03 LAMPBLACK            0  #16 LANDINGS            82
#02 LAMBDA               741  *01 LAMPEX               0  *01 LANDISMAN            0
 01 LAMBE                  0  *01 LAMPHOUSE            0  *01 LANDLINE             0
#02 LAMBERT               22  *01 LAMPING              0  *01 LANDLINES            0
*01 LAMBERTIAL             0  *01 LAMPLESS             0  *01 LANDLORD             0
*01 LAMBERTIAN            47  *03 LAMPLIGHT            0  #16 LANDMARK            17
*02 LAMBERTON              0  *01 LAMPORTS             0  #03 LANDMARKS           11
*01 LAMBERTS               5  *03 LAMPREY              7  *03 LANDMASS             9
*06 LAMBLIA                2  #03 LAMPS              119  *03 LANDMASSES           4
*01 LAMBRA                 0  *01 LAMPSMK              0  *03 LANDMINE             0
*03 LAMBS                  0  *01 LAMPSON              0  *03 LANDMINES            0
*01 LAMBWE                 0  *01 LAMS                 1  *01 LANDOLT              2
*01 LAMC                   0  *01 LAMSON               0  *01 LANDOUT              0
*01 LAMDA                 16  *01 LAMSTRESS            0  *01 LANDROCK             0
*01 LAMDBA                 2  *01 LAMTRAC              0  *01 LANDROLL             0
#02 LAME                  11  *01 LAMV                 0  *06 LANDROVER            0
#03 LAMELLA                4  #06 LAN                120  *06 LANDROVERS           0
*03 LAMELLAE              16  *01 LANA                 0  #16 LANDS               55
*02 LAMELLAR              22  *06 LANATUM              0  #02 LANDSAT            659
*01 LAMELLATED             0  *06 LANCASTER            3  *01 LANDSBERG            0
*01 LAMENTED               0  #06 LANCE                8  *01 LANDSBERGS           0
*01 LAMENTS                0  *01 LANCEJET             0  #03 LANDSCAPE          141
*01 LAMERE                 0  *01 LANCEOLA             0  *02 LANDSCAPED           0
*01 LAMILLOY               0  *06 LANCEOLATA           0  *06 LANDSCAPES          63
#03 LAMINA                98  *06 LANCER               0  *01 LANDSCAPING          0
*01 LAMINAC                0  *01 LANCET               5  *01 LANDSIDE            10
*01 LAMINAE               30  *02 LANCHESTER          10  *03 LANDSLIDE           19
#02 LAMINAR             1556  *01 LANCHESTERS          2  #03 LANDSLIDES          46
*03 LAMINARIA              0  *01 LANCORTEX            0  *01 LANDSMAN             0
*02 LAMINARISATION         0  *01 LANCZOS            104  *06 LANDSPACE            0
*01 LAMINARITIES           0  #03 LAND              1594   01 LANDSTREET           0
*01 LAMINARITY             1  *01 LANDA                0   01 LANDSTREETS          0
*01 LAMINARIZATION         5  *01 LANDAHL              0  *01 LANDURIGE            0
*02 LAMINARIZED            2  *01 LANDAHLS             0  *01 LANDWARD             0
*01 LAMINARIZING           0  *01 LANDAN               0  *01 LANDWEBER            0
*01 LAMINARY               0  #02 LANDAU              67  *06 LANE                45
*01 LAMINAS                3  *01 LANDBASE             0  *01 LANEPORT             0
*06 LAMINATE             462  *02 LANDBASED            0  #03 LANES               21
#02 LAMINATED            423  *01 LANDBIRDS            0  *01 LANET                0
#03 LAMINATES            760  *01 LANDBORNE            0  *01 LANFZ                0
*06 LAMINATING            11  *01 LANDE                0  *01 LANG                 0
*03 LAMINATION            66  *01 LANDECKER            0  *01 LANGAR               0
#03 LAMINATIONS           17  *01 LANDED              19  *06 LANGAT               0
*02 LAMINATIVE             0  *02 LANDEL               0  *01 LANGBERG             2
*06 LAMINATOR              0  *01 LANDEM               0  *01 LANGE                5
*03 LAMINECTOMIES          0  #16 LANDER             367  *01 LANGENSCHEIDTS       0
```

Figure A-2

APPENDIX B: GRAMMAR RULES FOR IDENTIFYING PHRASES

The machine phrase selection grammar and the word categories that are recorded in, and looked up in, the NASA Recognition Dictionary (NRD) are part of our first computer system for selecting phrases from text. The grammar rules are expressed as three columns of two digits each, where:

Column 1 identifies the grammar entrance point,
Column 2 identifies the word category, and
Column 3 identifies the exit point, which is also the next grammar rule entrance point, or the end-of-the-phrase indicator.

When the NRD has provided a category for each word in a string of text such as a title or a sentence, the machine phrase selection is ready to begin. The category of the first word is examined. If it is 01, the category of the next word is examined. As soon as a non-01 category is found, the program enters the grammar table looking for a rule (or a line of six digits) that begins with 40 (an arbitrarily chosen number used to identify the point of entry) and has the desired word category in column 2. For example, take the following sentence in which the word categories are indicated in parentheses:

CONTROLLED (02) SUBSTANCES (03) ARE (01) DANGEROUS (01).

The program begins by looking for the rules that start with 40 02 and finds:

```
40 02 41
40 02 42
```

The category of the next word in the example is 03. The program looks for rules beginning with the digits in column 3 followed by the category of the next word, i.e., 41 03 and 42 03. It finds:

```
41 03 41
41 03 80
```

but no rule beginning with 42 followed by 03. The path, or grammar rule selection process is much like going through a maze. When the path branches and there is more than one way to go, only one way will allow the program and phrase selection process to proceed. Other paths dead-end and block the program. The dead-end path, in this case the grammar rule 40 02 42, does not allow the program to proceed to any existing rule and so it is dropped and the other rule (40 02 41) is used.

The entire string, so far, is now 40 02 41 03 and either 41 or 80 and 01, which is the category of the next word. The only rule that begins with either 41 or 80 and 01 is: 80 01 81. Discarding the 41, which blocked the selection process, the string now reads: 40 01 41 03 80 01 81.

81 (another arbitrarily chosen number) is used to identify a legitimate exit from the grammar; therefore the program has now completed its selection of a phrase. The words selected, except the Category 01 or "stopword" at the end, are printed out as an acceptable or "good" phrase. In this example, the program has selected a string of words with the categories 02 03 (with a trace format of AN, i.e., adjective noun) which stands for the phrase CONTROLLED SUBSTANCES.

If the grammar blocks word additions instead of permitting an exit at 81, the program looks at the category of the last word in the candidate phrase string. All phrases must end with a word from one of the noun categories. If the last word of the candidate phrase string is of the wrong category, that word is dropped and the string is checked again. This process is repeated until an acceptable phrase format is reached or until all words have been dropped from the string.

MAPS Grammar

The MAPS grammar follows. It shows three 2-digit columns indicating first, the grammar entry point; second, the word category; and third, the exit or next entry point. 40 marks a start; 81 marks a stop.

40	02	41
40	02	42
40	03	41
40	03	80
40	06	41
40	14	41
40	14	46
40	17	44
40	17	45
40	19	41
40	19	46
40	19	80
40	20	47
40	21	43
40	22	41
40	22	46
41	02	41
41	02	42
41	03	41
41	03	80
41	06	41
41	06	80
41	08	71
41	14	41
41	16	41
41	16	80
*41	17	44
41	17	45
41	19	41
*41	19	80
41	20	41
41	22	41
*41	22	80
42	11	98
43	19	41
43	19	80
44	18	45
45	03	41
45	03	80
45	06	41
45	06	80
46	21	80
47	21	41
71	02	41
71	03	41
71	03	80
71	06	41
71	06	80
71	16	41
*71	16	80
*71	22	41
*71	22	80
80	01	81
*80	11	81
98	02	41
98	02	42
*98	08	41

Grammar rules on NASA's original list (dated 27 November 1984) have no asterisk (*). The rules flagged with an * were added after we had had some experience with MAPS. These new rules enabled MAPS to generate phrases and MAI to suggest terms that would otherwise have been lost.

APPENDIX C: LIST OF STOPWORDS USED WITH ACCESS-2

ABOUT
ABOVE
ACCOUNT
ACHIEVED
ACROSS
ADDITIONAL
AFTER
ALLOW
ALLOWS
ALONG
ALSO
ALTHOUGH
AMONG
AN
ANY
APPROPRIATE
APPROXIMATELY
ARBITRARY
ARE
AROUND
AS
ASPECTS
ASSOCIATED
ASSUMED
AVAILABLE
BASIS
BECAUSE
BEEN
BEING
BEST
BETTER
BOTH
BUT
CAN
CARRIED
CAUSED
CERTAIN
CHARACTERIZED
COMPARED
COMPLETE
CONSIDERATION
CONSIDERED
CONSISTS
CONTAINING
CONTAINS
CONVENTIONAL
CORRESPONDING
COULD
DEFINED
DEMONSTRATE

DEMONSTRATED
DESCRIBE
DESCRIBED
DESCRIBES
DESIGNED
DETAILED
DETERMINE
DETERMINED
DETERMINING
DEVELOP
DEVELOPED
DIFFERENT
DIRECTLY
DISCUSSED
DOES
DUE
DURING
E.G
EACH
EFFICIENT
EFFORTS
EITHER
EMPHASIS
EMPLOYED
ESPECIALLY
ESTABLISHED
EVALUATE
EVALUATED
EXAMINED
EXAMPLE
EXAMPLES
EXISTING
EXPECTED
EXPERIMENTALLY
FEW
FOUND
FULLY
FUNDAMENTAL
FURTHER
GIVEN
GOOD
GREATER
HAD
HAS
HAVE
HAVING
HERE
HOW
HOWEVER
IDENTIFIED

I.E
IF
IMPLEMENTATION
IMPORTANCE
IMPORTANT
IMPROVE
INCLUDE
INCLUDED
INCLUDES
INCLUDING
INCREASE
INCREASED
INCREASES
INDICATE
INDIVIDUAL
INTEREST
INTO
INTRODUCED
INVESTIGATE
INVESTIGATED
INVOLVED
INVOLVING
IS
ISSUES
IT
ITS
KNOWN
LESS
MADE
MAJOR
MAKE
MAY
MEANS
MORE
MOST
MUCH
MUST
NECESSARY
NEED
NEEDED
NOT
OBJECTIVE
OBSERVED
OBTAIN
OBTAINED
OCCUR
OTHER
OUR
OVERALL
PART

PARTICULAR
PAST
PERFORMED
POSSIBLE
PREDICT
PREDICTED
PRELIMINARY
PRESENCE
PRESENT
PRESENTED
PRESENTS
PREVIOUS
PREVIOUSLY
PRODUCE
PRODUCED
PROPOSED
PROVIDE
PROVIDED
PROVIDES
PROVIDING
RECENT
RELATED
RELATIVELY
REPORTED
REQUIRED
REQUIRES
RESPECT
RESULT
RESULTING
RESULTS
REVIEWED
RTOP
SAME
SELECTED
SEVERAL
SHOULD
SHOW
SHOWED
SHOWN
SHOWS
SIGNIFICANT
SIGNIFICANTLY
SINCE
SOME
STATUS
STUDIED
STUDIES
STUDY
SUB
SUCH

SUGGESTED
SUITABLE
SUMMARY
TAKEN
TESTED
THAN
THAT
THEIR
THEM
THEN
THERE
THESE
THEY
THIS
THOSE
THROUGH
THUS
TOGETHER
TOWARD
TYPES
TYPICAL
UNDERSTANDING
UNIQUE
UP
UPON
USED
USEFUL
USES
USING
VARIETY
VARIOUS
VERSION
VIA
WAS
WE
WERE
WHEN
WHERE
WHICH
WHILE
WHOSE
WILL
WITH
WITHIN
WITHOUT
WOULD
YEARS

APPENDIX D: SPECIFICATIONS FOR A SINGLE-WORD TERM ASSESSMENT TOOL

Many of the single-word Thesaurus terms that appear in MAI output lists are recognized by indexers as inappropriate or unnecessary. The procedure described below will provide a tool for the assessment of KB entries for these single-word terms.

I. Creation of an INPUT file

A. Identify and capture all single-word thesaurus terms and all terms with a single word plus a parenthetical gloss.

B. Creat a file that includes thesaurus-term/text-word pairs. "Text-words" are generated for each Thesaurus term (from step A) by:

1) dropping parenthetical glosses, if any, and duplicating the single-word "term"

2) adding variants of the words obtained in step 1. (That is, "s" is added to words that do not end in "s" and "s" is removed from words that do end in "s".)

Individual elements of the INPUT file consist of a single, unique text-word and its associated thesaurus term.

Example:

Text-word	Thes.-term
plasmas	PLASMAS (PHYSICS)
plasma	PLASMAS (PHYSICS)
spectrometers	SPECTROMETERS
spectrometer	SPECTROMETERS
tempering	TEMPERING
temperings	TEMPERING

C. Compare "text-word" list with KB keys and delete any INPUT file elements with text-words having "00" as the KB posting term.

II. Determination of "usage/occurrence" ratios

A. General Procedure

For each element of the INPUT file determine the number of times that the text-word occurs in either the title or abstract of the database records for which MAI is used. Records with occurrences of the text-word in both fields are counted only once.

Then determine the number of these records that also contain the thesaurus term in the subject term field (MJS and MNS).

Express the ratio as a percentage.

B. Programming Considerations

- o This tool is to be produced only once, therefore the development of a formal, coherent program is not necessary.

- o The records counts required for this procedure can currently be determined using RECON Search and Combine operations as follows:

	Set#	
Select utp/plasmas	1	= text-word in title
Select ax/plasmas	2	= text-word in abstract
Select st/PLASMAS (PHYSICS)	3	= Thesaurus term
Combine 3and(1or2)	4	= Boolean operation

  The use of existing RECON processing capabilities should be considered.

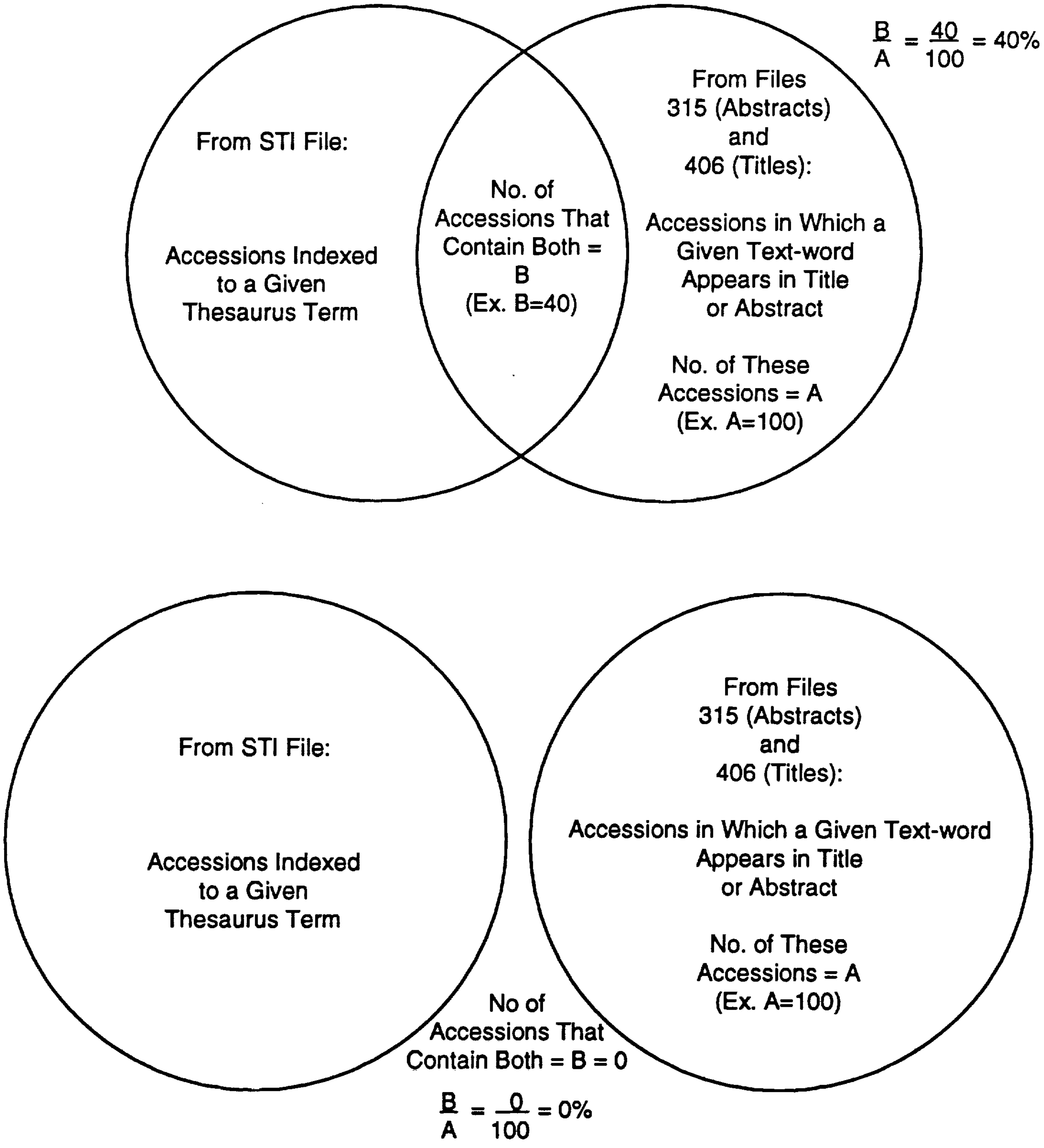
- o If a RECON-based procedure is not used it may be necessary to capture record accession numbers along with the quantities required in part A. (See the following section.)

C. Suggested "Non-RECON" Procedure

Access the inverted files that contain the postings of the linear files that contain (1) the text of the abstracts (File 315), (2) the text of the titles (File 406), and (3) the thesaurus terms (STI file).

Capture the accession numbers of the records that contain the text-word in either the abstract or title. Count the number of unique accessions and let this count equal set "A". See Figure D-1.

Ratio Wanted = B/A, Use for Sort of Elements (Word Pairs)



$\frac{B}{A} = \frac{40}{100} = 40\%$

From STI File:

Accessions Indexed
to a Given
Thesaurus Term

No. of
Accessions That
Contain Both =
B
(Ex. B=40)

From Files
315 (Abstracts)
and
406 (Titles):

Accessions in Which a
Given Text-word
Appears in Title
or Abstract

No. of These
Accessions = A
(Ex. A=100)

From STI File:

Accessions Indexed
to a Given
Thesaurus Term

From Files
315 (Abstracts)
and
406 (Titles):

Accessions in Which a Given Text-word
Appears in Title
or Abstract

No. of These
Accessions = A
(Ex. A=100)

No of
Accessions That
Contain Both = B = 0

$\frac{B}{A} = \frac{0}{100} = 0\%$

Figure D-1

Capture the accession numbers of the records that contain the corresponding thesaurus terms in the major or minor subject term field.

Determine which accession numbers contain both the text-word and the thesaurus term. Count these accessions and let this count equal set "B".

Express the ratio B/A as a percentage rounded to the nearest whole number. Show percentages less than 1 as 0%. Assign this percentage to the text-word.

Example:

For the INPUT element

"plasma --> PLASMAS (PHYSICS)"

The text-word "plasma" occurs in 55 accessions in file 315 (abstract file) and 50 accessions in file 406 (title file). A total of 100 unique accessions (set A) are captured from both files (5 accessions being common to both files).

The corresponding thesaurus term PLASMA (PHYSICS) occurs in 66 accessions in the STI (thesaurus term postings file). The 66 accessions are captured.

The two accession lists are compared and the number of accessions which both lists have in common is determined (in this case 33 items - set B)

B/A is 33/100 or 33%.

III. Output Format

The output will include percentages, text-words, thesaurus terms, and the values of "B" and "A". The output will be sorted by percentage values (lowest to highest) with a secondary alpha ordering of the text-words.

Example:

%	Text-word	Thes.-word	B/A
0%	temperings	TEMPERING	1/200
33%	plasma	PLASMAS (PHYSICS)	33/100
33%	spectrometer	SPECTROMETERS	66/200
40%	plasmas	PLASMAS (PHYSICS)	40/100

APPENDIX E: ONLINE MAINTENANCE COMMANDS

The NLD's online maintenance system provides a set of general purpose commands that allow maintenance personnel to process input from any of the sources described in the section above. There is a separate set of commands for each of the NLD data files. The chart below indicates the capabilities provided by the maintenance system, along with the command used for each of the NLD data files to carry out that function. A "Request to Run" form is required by Computer Operations each time any of the following commands are used: VALSETUP, NASAVAL, NASALOAD, NASAPRNT, NASANVRT, NASAUNLD. (This is also true for the corresponding commands for DOE and DTIC, listed below.)

Maintenance capability:	Maintenance System Command:		
		DOE and DTIC	
	MAI KB	Subj. Switching Files	
Creates authority files	VALSETUP	DOEVSAM	DTICVSAAM
Validates data file	NASAVAL	DOEVAL	DTICVAL
Enters update transactions	*	DOEUPDT	DTICUPDT
Loads update transactions	NASALOAD	DOELOAD	DTICLOAD
Prints file, alpha by key	NASAPRNT	DOEPRNT	DTICPRNT
Prints file, alpha by postings	NASANVRT	DOENVRT	DTICNVRT
Displays 10 records online	NASAFIND	DOEFIND	DTICFIND
Provides NLD translations online	NEWACC	DOEACC (mode 2)	DTICACC (mode 2)

*The former maintenance command NASAUPDT has been replaced with the capability of going directly into a transaction dataset and typing the desired entry. These transactions are then copied into the modification dataset, which is editable online. When the command NASALOAD is given to load the modifications into the MAI KB, the transactions in the dataset are checked for the correct format and for a valid posting term or terms. Only those entries that are judged to be correct will be loaded into the MAI KB. The original NASAUPDT program rejected formatting errors at the time of data entry. The new method rejects any incorrect entry at the time of loading them into the master file, i.e., the MAI KB.

The MAI commands are described more fully in the pages that follow. The DTIC commands are described in NASA-CR-3838 and DOE commands are the same except that they act on the DOE file(s).

VALSETUP. This command creates two authority files for NASA Thesaurus terms from the online Thesaurus files:

o A sequential file of NASA posting terms and Use references. This file is used by the validation routine to check that there is an entry, that is, a key to a record, for every NASA Thesaurus term and Use reference.

- A VSAM file of NASA posting terms only: "NLD.THES.TERMS'. The VSAM file is used by NASAVAL to verify that all posting terms that appear in the posting term field of existing entries in the KB are valid NASA Thesaurus terms.

As the VSAM file is being created, each term is checked against the KB ('NLD.NASA.MASTER') to determine if it should be marked as an array term and to add an "at sign" (@) to the term if it is. Most array terms are given a null translation (00). A few are MAI-suggested but flagged with @ to alert the indexer to the fact that a more specific term is to be preferred. See the section on "Problems" following "Text Processing with Access-1."

NASAVAL. This command initiates comparisons between the entries in the MAI KB and the NASA Thesaurus authority files. NASAVAL checks:

- Every term appearing in the posting term field against the NASA authority file. If a KB posting term does not appear in the Thesaurus authority file, an error message is generated.

- Every posting term and Use reference appearing in the NASA Thesaurus authority file against the MAI KB keys. Each of these terms should appear as a key in the KB, and an error message is generated if it does not.

- Every posting term in the NASA Thesaurus authority file against the KB posting terms. If a Thesaurus posting term does not also appear as a KB posting term, an error message is generated.

These error messages highlight the additions, modifications, and deletions required in the KB.

NASALOAD. This command does a series of edit and validation checks. Those transactions that pass the checks are loaded into the KB. Those transactions that do not pass the checks are written out on an error list and returned to the modifications dataset. The edit checks reject entries that have any of the following:

Invalid Characters. Valid characters are: A-Z, 0-9, +, ?, >, &, ', $, (,), ;, ., %, *, /, @, -, ,(a comma), or blank.

Invalid Characters in the Logic Code. The logic code is recorded in the first 3 columns of the KB entry. Valid codes are: DEL; a zero and two blanks, or one of the following alphas and two blanks: C, E, L, or T.

Logic Code Too Long. More than 3 characters or blanks appear before the key in the transaction.

Logic Code All Blanks. Three blanks appear before the key in the transaction.

No posting term. Nothing appears following the $ after the key.

Too Many $'s. More than one $ has been entered in the transaction. (In the online entering of a transaction, only one $ is used. It separates the key from the posting term(s). In writing out the transaction on paper, a $ is placed both before and after the key to indicate the end of the logic code field and the key field.)

Invalid Format. The transaction does not conform to one of the following formats:

```
Logic code$Word1;Word2$Posting term(s) or 00 or *
Logic code$Word;00$Posting term(s) or 00 or *
DEL$(Key of record to be deleted)
```

Transaction Posting Terms Not Found in the Thesaurus. The posting term may be invalid or may have been removed from the thesaurus.

The person doing the maintenance corrects the rejected transactions in the modification dataset and re-executes the NASALOAD command. Following the execution of the NASALOAD command, any printout generated is examined by NLD personnel. This is to:

- See whether or not the job has run satisfactorily. (The Job Control Language return codes (RC) should all equal zeros.)
- See whether or not there are any errors listed that must be corrected.
- Record the number of changes and additions to, and deletions from, the KB. These figures are accumulated for reports.
- See whether or not any KB entry that has been changed needs any further adjustment. For example:
    - if a continuation entry is changed to a posting term, the continuation entry must be replaced and a new key ending in 999 must be entered going to the NASA posting term(s).
    - if the new entry has a less inclusive translation than the old entry, the old entry may be preferred. This correction could be avoided by checking the KB before making the entry that now needs to be changed again, but it is generally less time-consuming to check the printout after the fact than the database beforehand.

NASAPRNT. This command generates a print of the KB sorted alphabetically by keys. A sample page of the revised KB, that is, without logic codes, is shown in Figure E-1.

NASANVRT. This command generates a print of the KB sorted alphabetically by posting terms. In order to readily locate a particular posting term in the KB, it is necessary to have a print of the KB sorted alphabetically by posting term. Entries with multiple posting terms are listed once for each posting term. A sample page is shown in Figure E-2.

```
  08/01/92              NASA LEXICAL DICTIONARY                        PAGE    1
(HG,CD)TE;999                                        CADMIUM TELLURIDES,
                                                       MERCURY TELLURIDES
&;999                                                OO
/U/;999                                              OO
A;AND                                                *
A;AND;B                                              *
A;AND;B;STARS                                        A STARS,
                                                       B STARS
A;AND;B-TYPE                                         A STARS,
                                                       B STARS
A;AND;F                                              *
A;AND;F;MAIN                                         *
A;AND;F;MAIN;SEQUENCE                                A STARS,
                                                       F STARS,
                                                       MAIN SEQUENCE STARS
A;AND;F;STARS                                        A STARS,
                                                       F STARS
A;AND;F;TYPE                                         A STARS,
                                                       F STARS
A;AND;F-TYPE                                         A STARS,
                                                       F STARS
A;BAND                                               OO
A;GIANT                                              A STARS,
                                                       GIANT STARS
A;GIANTS                                             A STARS,
                                                       GIANT STARS
A;STAR                                               OO
A;STARS                                              A STARS
A;SUPERGIANT                                         A STARS,
                                                       SUPERGIANT STARS
A;SUPERGIANTS                                        A STARS,
                                                       SUPERGIANT STARS
A;TYPE                                               *
A;TYPE;SHELL                                         *
A;TYPE;SHELL;STAR                                    A STARS
A;TYPE;SHELL;STARS                                   A STARS
A;TYPE;STAR                                          A STARS
A;TYPE;STARS                                         A STARS
A;1367                                               GALACTIC CLUSTERS
A;999                                                OO
A-;AND                                               *
A-;AND;B-STARS                                       A STARS,
                                                       B STARS
A-;999                                               OO
A-ALKYLACRYLATE;POLYMERS                             ACRYLIC RESINS
A-ALKYLACRYLATE;999                                  ACRYLATES,
                                                       ALKYL COMPOUNDS
A-B;BINARY                                           BINARY MIXTURES
A-B;999                                              OO
A-BAND;999                                           OO
A-C;999                                              ALTERNATING CURRENT
A-F;STARS                                            A STARS,
                                                       F STARS
A-I;TECHNOLOGY                                       ARTIFICIAL INTELLIGENCE
A-M/AGC;999                                          AUTOMATIC GAIN CONTROL
A-SHELL;STAR                                         A STARS
A-SHELL;STARS                                        A STARS
A-SI:H;FILMS                                         AMORPHOUS SILICON,
```

Figure E-1

T	COMMERCIAL;LAUNCH;VEHICLES	SPACE COMMERCIALIZATION, LAUNCH VEHICLES
T	COMMERCIAL;LAUNCH;VEHICLE	SPACE COMMERCIALIZATION, LAUNCH VEHICLES
T	EARTH;TO;ORBIT;VEHICLE	LAUNCH VEHICLES
T	EARTH;TO;ORBIT;VEHICLES	LAUNCH VEHICLES
T	EARTH-TO-ORBIT;LAUNCH;VEHICLE	LAUNCH VEHICLES
T	EARTH-TO-ORBIT;LAUNCH;VEHICLES	LAUNCH VEHICLES
T	EARTH-TO-ORBIT;VEHICLES	LAUNCH VEHICLES
T	BOOSTER;VEHICLE	LAUNCH VEHICLES
T	BOOSTER;VEHICLES	LAUNCH VEHICLES
T	AIR-BREATHING;LAUNCH;VEHICLE	AIR BREATHING BOOSTERS, LAUNCH VEHICLES
T	AIR;BREATHING;LAUNCH;VEHICLES	AIR BREATHING BOOSTERS, LAUNCH VEHICLES
T	AIR-BREATHING;LAUNCH;VEHICLES	AIR BREATHING BOOSTERS, LAUNCH VEHICLES
T	CARRIER;ROCKET	LAUNCH VEHICLES
T	CARRIER;ROCKETS	LAUNCH VEHICLES
T	AIR;BREATHING;LAUNCH;VEHICLE	AIR BREATHING BOOSTERS, LAUNCH VEHICLES
T	SOVIET;LAUNCH;VEHICLES	LAUNCH VEHICLES, SOVIET SPACECRAFT
T	SOVIET;LAUNCH;VEHICLE	LAUNCH VEHICLES, SOVIET SPACECRAFT
T	LAUNCH;WINDOW	LAUNCH WINDOWS
T	LAUNCH;WINDOWS	LAUNCH WINDOWS
T	LAUNCH;TIME	LAUNCH WINDOWS
T	LAUNCH;OR;LANDING;WINDOW	LAUNCH WINDOWS, SPACECRAFT LANDING, WINDOWS (INTERVALS)
C	LAUNCHER;OO	LAUNCHERS
T	ELECTROMAGNETIC;LAUNCHERS	ELECTROMAGNETIC PROPULSION, LAUNCHERS
T	BOX;LAUNCHER	LAUNCHERS
T	BOX;LAUNCHERS	LAUNCHERS
E	LAUNCHERS;OO	LAUNCHERS
T	LAUNCHING;DEVICE	LAUNCHERS
T	LAUNCHING;DEVICES	LAUNCHERS
T	LAUNCH;TUBES	LAUNCHERS
T	LAUNCH;MODES	LAUNCHING
T	LAUNCH;OO	LAUNCHING
T	LIFT-OFF;OO	LAUNCHING
T	LIFT;OFF	LAUNCHING
C	GROUND/LAUNCH;OO	LAUNCHING
T	LAUNCHING;BASE	LAUNCHING BASES
T	LAUNCH;COMPLEX;OO	LAUNCHING BASES
T	LAUNCH;FACILITIES	LAUNCHING BASES
T	LAUNCH;FACILITY	LAUNCHING BASES
T	LAUNCH;COMPLEXES	LAUNCHING BASES
T	LAUNCH;CONTROL;CENTER	LAUNCHING BASES
T	LAUNCH;CONTROL;FACILITIES	LAUNCHING BASES
T	LAUNCH;CONTROL;FACILITY	LAUNCHING BASES
T	LAUNCHING;BASES	LAUNCHING BASES
T	LAUNCHING;COMPLEXES	LAUNCHING BASES
T	LAUNCHING;FACILITIES	LAUNCHING BASES
T	LAUNCHING;FACILITY	LAUNCHING BASES
T	LAUNCHER;COMPLEXES	LAUNCHING BASES

NASAFIND. This command searches the KB for a specified key or first word in a key, and displays ten sequential KB records, beginning with the key or word requested, if it exists. If the requested key or word is not found, the program will locate the sequential position in which it should occur and display the next ten records. NASAFIND is a quick way of displaying the KB, which is a VSAM dataset. A more flexible way of displaying the records in a VSAM dataset is to type out:

```
PRINT IDS(dataset.name.here) CHAR COUNT(xx) FK('key')
```

NEWACC. This command processes a maximum of one line of text or a minimum of one word through the Access-2 program. It provides, on the terminal screen, the full or partial translation of the input material, if any translation into NASA terms is available through the KB. Otherwise the program returns the message:

```
UNABLE TO IDENTIFY
```

The command can be used to test how MAI will translate words, phrases, or groups of phrases. A sample of NEWACC output is illustrated in Figure E-3.

```
                                                                                  4/92-12:52:16-PM

           ENTER A  /*  TO TERMINATE PROCESSING

                        PLEASE ENTER PHRASE

> abundances in chemically peculiar and normal A-type stars

ABUNDANCES;999 .USE. ABUNDANCE

IN;999 .USE. 00

CHEMICALLY;999 .USE. 00

PECULIAR;999 .USE. 00

AND;999 .USE. 00

NORMAL;A;STARS .USE. A STARS

A;TYPE;STARS .USE. A STARS

PLEASE ENTER PHRASE

>
```

Figure E-3

REFERENCES

Artandi, Susan: Machine Indexing: Linguistic and Semiotic Implications. JASIS, vol. 27, no. 4, 1976, pp. 235-239.

Jones, Leslie P., Gassie, Edward W., Jr., Radhakrishnan, Sridhar, INDEX: The Statistical Basis for an Automatic Conceptual Phrase-Indexing System. JASIS, vol. 41, no. 2, Mar. 1990, pp. 87-97.

Klingbiel, Paul H., Phrase Structure Rewrite Systems in Information Retrieval. Information Processing and Management, vol. 21, no. 2, 1985, pp. 113-126,

Lancaster, Frederick W.: Indexing and Abstracting in Theory and Practice. Univ. of Ill., Champaign, 1991, pp. 60-85.

Lustig, G., Knorz, G.: Pilotanwendung von Automatischen Indexing und Verbesserten Retrievalverfahren mit der Datenbank PHYS (AIR/PHYS Pilot Application Project: Pilot Application of Automatic Indexing and Improved Retrieval Methods Using the PHYS Data Base), Fachinformationszentrum, Energie Physik Mathematik GmbH, Karlsruhe, Federal Republic of Germany, 1986, pp. 1-30.

Silvester, June P., Newton, Roxanne, and Klingbiel, Paul H., An Operational System for Subject Switching Between Controlled Vocabularies: A Computational Linguistics Approach. NASA-CR-3838, 1984.

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
March 1993

3. REPORT TYPE AND DATES COVERED
Contractor Report

4. TITLE AND SUBTITLE
Machine Aided Indexing from Natural Language Text

5. FUNDING NUMBERS

6. AUTHOR(S)
June P. Silvester, Michael T. Genuardi and Paul H. Klingbiel

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
NASA Center for AeroSpace Information
Linthicum Heights, MD 21090-2934

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
National Aeronautics and Space Administration
Washington, DC 20546

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
NASA-CR-4512

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Unclassified - Unlimited
Subject Category - 82

12b. DISTRIBUTION CODE

13. ABSTRACT *(maximum 200 words)*

The NASA Lexical Dictionary (NLD) Machine Aided Indexing (MAI) system was designed to (1) reuse the indexing of the Defense Technical Information Center (DTIC); (2) reuse the indexing of the Department of Energy (DOE); and reduce the time required for original indexing. This was done by automatically generating appropriate NASA thesaurus terms from either the other agency's index terms, or, for original indexing, from document titles and abstracts. The NASA STI Program staff devised two different ways to generate thesaurus terms from text. The first group of programs identified noun phrases by a parsing method that allowed for conjunctions and certain prepositions, on the assumption that indexable concepts are found in such phrases. Results were not always satisfactory, and it was noted that indexable concepts often occurred outside of noun phrases. The first method also proved to be too slow for the ultimate goal of interactive (online) MAI. The second group of programs used the knowledge base (KB), word proximity, and frequency of word and phrase occurrence to identify indexable concepts. Both methods are described and illustrated. Online MAI has been achieved, as well as several spinoff benefits, which are also described.

14. SUBJECT TERMS
computer techniques, dictionaries, information retrieval, information systems, terminology, thesauri

15. NUMBER OF PAGES
83

16. PRICE CODE
A05

17. SECURITY CLASSIFICATION OF REPORT UNCLASS

18. SECURITY CLASSIFICATION OF THIS PAGE UNCLASS

19.SECURITY CLASSIFICATION OF ABSTRACT UNCLASS

20. LIMITATION OF ABSTRACT
UNLIMITED

Available from NASA Center for AeroSpace Information
800 Elkridge Landing Road
Linthicum Heights, MD 21090
(301) 621-0390-2934