NASA Contractor Report 4518

NASA's Online Machine Aided Indexing System

Final 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
Office of Management
Scientific and Technical
Information Program
1993

TABLE OF CONTENTS

FIGURES

PREFACE

This report describes the NASA Lexical Dictionary (NLD), a machine aided indexing system used at the National Aeronautics and Space Administration's Center for AeroSpace Information (CASI). The NLD was developed for NASA's Scientific and Technical Information Program. CASI's current NLD operates as part of its online Input Processing System, uses a program called Access-2, and a knowledge base of nearly 115,000 entries or rules. The NLD system automatically suggests a set of candidate terms from NASA's controlled vocabulary for any designated natural language text input. Development of the system was begun under contract NASw-3330 by the Planning Research Corporation/Government Information Systems and was continued by RMS Associates under contracts NASw-4070 and NASw-4584.

We would like to acknowledge the help given to our project by the people at the Defense Technical Information Center who shared their expertise, their programs, their files, and their controlled vocabulary with us when we began the NLD project. Without their cooperation the construction of the NLD would have been more difficult and more costly. We also acknowledge the encouragement and opportunities provided by Herman Miles of Planning Research Corporation and John Wilson of NASA Headquarters Scientific and Technical Information Program. The machine aided indexing project was undertaken because of their foresight.

ABSTRACT

This report describes the NASA Lexical Dictionary, a machine aided indexing system used online at the National Aeronautics and Space Administration's Center for AeroSpace Information (CASI). This system automatically suggests a set of candidate terms from NASA's controlled vocabulary for any designated natural language text input. The system is comprised of a text processor that is based on the computational, non-syntactic analysis of input text, and an extensive 'knowledge base' that serves to recognize and translate text-extracted concepts.

The structure and function of the various NLD system components are described in detail. Methods used for the development of the knowledge base are discussed. Particular attention is given to a statistically-based text analysis program that provides the knowledge base developer with a list of concept-specific phrases extracted from large textual corpora.

Production and quality benefits resulting from the integration of machine aided indexing at CASI are discussed along with a number of secondary applications of NLD-derived systems including on-line spell checking and machine aided lexicography.

BENEFITS AND APPLICATIONS

The NASA machine aided indexing system, known as the NASA Lexical Dictionary (NLD), is a proven time-saver. As an aid to human indexers, it generates authorized, NASA index terms from any given input. Usually this input consists of document titles and abstracts, but it may include index terms assigned by another organization, or any computer-identifiable text. The use of machine aided indexing and a controlled vocabulary also reduces retrieval problems caused by synonymous words and phrases. The preferred words and phrases as listed in the Thesaurus are automatically presented to the indexer for acceptance or deletion.

Indexers save time with the online NLD system because

- terms suggested by machine aided indexing are automatically selected, displayed online in the correct format (with the authorized word endings - such as tion or ing), accurately spelled, and machine-readable;
- indexer research time is reduced because natural language words and phrases and their technical language/Thesaurus equivalents are researched before they are added to the NLD database, thereby presenting the indexer with expert advice;
- appropriate, unfamiliar, technical terms may be suggested which the indexer would omit without a prompt from machine aided indexing; and
- terms suggested by machine aided indexing function as a check-list of indexable concepts and increase the consistency of indexing.

Another time-saver at the NASA Center for AeroSpace Information is the electronic Input Processing System (IPS) of which online machine aided indexing is an integral part. The online system saves time for the indexer in editing MAI output because

- the suggested NASA terms are editable with very few keystrokes:
  - acceptable major terms require no action on the part of the indexer to accept the term;
  - acceptable minor terms require two keystrokes to specify that it is a minor, rather than a major term;
  - unwanted terms are eliminated with a single keystroke;
  - additional terms (up to 42 characters in length) can be transferred from an online thesaurus display with four or five keystrokes.

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 primarily in cases where 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 6 seconds for a title plus a 150-250 word abstract, receives a group of thesaurus terms for consideration. The number varies according to the input, but averages from 7-12 terms. 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.

o IPS also allows already-written abstracts to be entered into the record electronically with a scanner. Original abstracts are composed online at a computer terminal.

Although the NLD was intended to be an aid for indexers, other uses were discovered, particularly applications that require identification of NASA Thesaurus terms from equivalent natural language words and phrases, or vice versa.

The Thesaurus lexicographer uses the NLD to identify Thesaurus terms that appear within the definitions of other Thesaurus terms. If these identified terms are also defined, they are printed in bold type. This provides a cross reference capability not possible without machine aided indexing.

The retrieval analysts benefit from machine aided indexing because it has increased the number of index terms assigned and therefore the number of access points for those documents.

Proofreaders benefit from the use of the NLD as a custom-built spell-checker. The knowledge base, which is discussed under SYSTEM COMPONENTS, has been expanded to enhance this use.

The NLD has also been used to generate sets of NASA Thesaurus terms for documents previously indexed only with Library of Congress' subject headings.

In another instance, the NLD was helpful in re-indexing more than 400,000 records in the NASA database. These were records that were indexed before the controlled vocabulary existed. (See Appendix C, "Automatic Re-indexing of NASA's Pre-thesaurus STI Records," for a description of this project.)

Potential applications, not yet explored, are as a front end for searching files in the online document retrieval system (currently RECON); as an indexing aid for machine readable full text; as a tool for integrating various thesauri; and as part of an automatic indexing system for generating controlled Thesaurus terms for historical collections.

The NASA NLD system is significant for its versatility. New applications continue to keep the system involved in experimentation. At the same time, the NLD 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 operating machine aided indexing (MAI) systems 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.

EVALUATION MEASURES

Machine aided indexing (MAI) was developed at the NASA Center for AeroSpace Information in a high-pressure, production environment. Measurements of its 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 per person has approximately doubled in the past 10 years.
- 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 NLD system, have been able to maintain and even improve indexing quality, and at the same time increase production. However, MAI is not the only change responsible for increased productivity. Input is now done at a computer terminal or with a scanner or electronically from magnetic tape instead of with pen and paper. This also speeds processing. 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 the term for the equipment described, while MAI may suggest the process.

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

It was determined in an early test with experienced indexers 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 who have not become thoroughly familiar with the variety of terms in NASA's controlled vocabulary. This currently amounts to about 40% of the indexing staff.

Match Rate

Another early measure of how well the MAI system performed was referred to as the match rate. Project Director Klingbiel originally used this to describe the percentage of machine selected words or phrases (semantic units) that either entirely or partly matched a key in the knowledge base. When that percentage reached the upper 90's, it lost its value as a measure of progress, and so it was redefined.

The match rate now refers to the percentage of MAI-suggested terms that the indexer elects to use. This measure which began at 23% now ranges from 40% to 60% - or an average of 50% - rising gradually as improvements are made to the system.

Capture Rate

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

Consistency Factor

In late 1989, we began to calculate the consistency (or quality) factor "q". This identifies the percentage of common terms "c" found in two lists of terms, one generated automatically and represented 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).

The following table shows the match rate, capture rate, and consistency factors calculated for 1987, 1988, and current estimated performance.

Year	Match Rate	Capture Rate	Consistency Factor
1987	32.4 %	36.9 %	20.8 %
1988	37.0	39.0	23.4
1993	50.0	50.0	33.3

The 1987 figures are for a sample of approximately 2,500 documents. The 1988 figures were based on a sample of 100 documents, and the 1993 figures are from a survey of the indexers currently using the system.

SYSTEM COMPONENTS

NASA's online machine aided indexing system has three components:

(1) an application program that
- o indicates the input text to be processed;
- o selects text strings from the specified text;
- o "calls" Access-2 and feeds those strings to it;
- o accepts and stores NASA terms from Access-2;
- o prints out various reports; and
- o for NASA's electronic Input Processing System (IPS), provides an online display of NASA terms.

The application programs differ for each use of machine aided indexing. The reports printed usually consist of a list of natural language words and phrases selected from the strings by Access-2 with their equivalent NASA Thesaurus terms, and a list of words not found in the knowledge base (see system component number 3).

(2) Access-2, a program that
- o accepts strings from the applications program;
- o identifies single words or word combinations from the strings;
- o looks up these words or combinations in the knowledge base (KB); and
- o sends to the application program the appropriate NASA Thesaurus terms and any other reports for output to the user.

Access-2 was written as a modular program that is "called" by any applications program.

(3) The knowledge base (KB), a data file that
- o contains records consisting of two fields:
  - Key; and
  - NASA Posting Term(s);
- o provides Thesaurus term equivalents for input natural language words or word combinations;
- o suppresses unwanted translations of natural language; and
- o directs the computer to look for word combinations of more than two words when they exist in the KB.

The KB is currently stored in a Virtual Storage Access Method (VSAM) file that contains nearly 115,000 records (or rules). It is described at greater length below.

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 KB pays off with high quality output and indexer acceptance.

The Key Field

The Key field is the part of the dataset that words or phrases from natural language text are matched against. These words or phrases that are selected from natural language text are semantic units, chosen not through syntactic parsing, but through proximity rules. (See the section below on text processing.) The key field of a record is unique and serves as the address to the record. The key may contain one or as many as seven words. The number of words that a key may contain is limited only by the length of the field - currently 80 characters. Semicolons 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 following the semicolon is null. This is marked with three nines, because NASA computers sort nines last.

Example: DIFFUSION;999

It is important to have the entry with the null value sort last because the computer looks at the multiword entries that begin with the same word alphabetically. The entry that has the null value 999 following the last semicolon is a default, to be used only when no other entry can be found. 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 toward the multi-word key. The key of the first entry consists of the first two words with a semicolon in between. Each continuation entry adds a semicolon and one more word to the key until the key contains the entire phrase. If more than five words are in a key, they must appear in text consecutively in order to translate.

The Posting Term Field

The Posting Term field of a record contains one of three things:

(1) the NASA term or terms that MAI will suggest for indexing the concept expressed by the word or phrase in the key;

(2) two zeros, which will "zero out" or suppress the return of any NASA Thesaurus terms; or

(3) an asterisk, which tells the computer to look for another word or for ";999" to find a key that will translate. Translations for semantic units to NASA Thesaurus terms occur when the phrase in the key is complete.

The posting term field may contain multiple Thesaurus terms. To save storage space (an early concern for system designers), terms are separated by commas, but no spaces. Spaces are preserved only between words in multiword terms. The length of the field is variable, but has been limited arbitrarily to 150 characters. Originally both the Key and Posting Term fields were longer, but the reduced length was found to be adequate to handle any entry in the KB. Leaving out spaces in the Posting Term field and shortening both fields also reduced the time that the computer was obliged to take in order to read all the recorded blanks in each record.

If the natural language word or phrase being searched for (the semantic unit) is not of indexable importance, the KB can delete the unit either by omitting it from the KB or by providing a posting of 00.

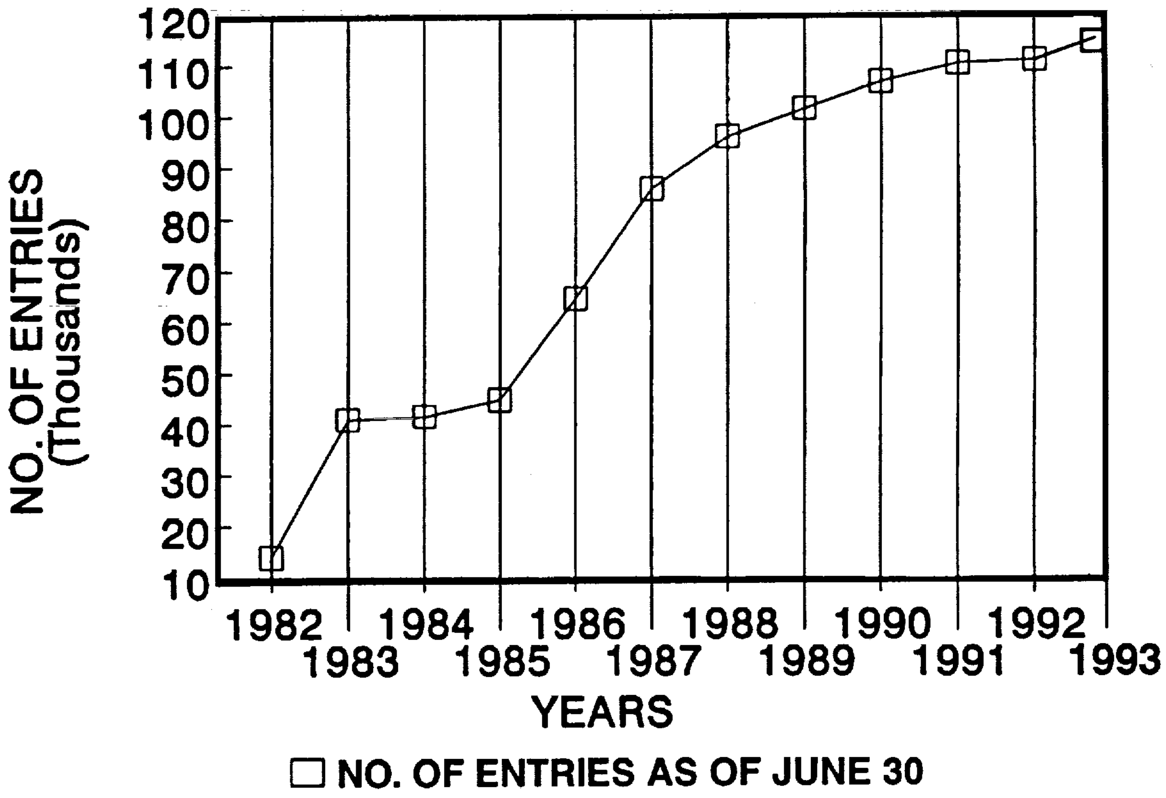
When a translation is wanted for the key of a continuation entry, a semicolon and three nines are added at the end of the key, to create a unique key. The appropriate translation is then entered in the posting term field of the unit that ends with 999.

KNOWLEDGE BASE DEVELOPMENT

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 worked much of the time on building another database which translates 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 use of the KBB Text Analysis Tool, which aids analysts in the selection of good quality entries for expanding the KB. The KB today has nearly 115,000 records, occupying 589 tracks, or about 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 a text-based MAI system such as NASA's, semantic analysis is dependent upon the content of the KB. Primary concerns with this component 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 originally limiting analysis to certain syntactic phrase forms. Changing the method of phrase selection from a syntactic process to a proximity search process eliminated the last problem, and the KB Text Analysis Tool reduced several of the others.

KNOWLEDGE BASE GROWTH



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	1993	115,400

1983-1984 SHOWS EFFECT OF REDUCTION IN RESOURCES
1991-1992 ALSO SHOWS EFFECT OF REDUCTION IN RESOURCES
FOR THIS PARTICULAR TASK

FIGURE 1

Content of 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 machine aided indexing system, the system designers must be concerned with the tradeoffs among

(1) the size of the KB file,
(2) the system's response time, and
(3) the level of complexity selected.

By level of complexity is meant the number of special rules the system must consider before it provides the thesaurus terms for indexer review. These rules might include, but need not be limited to:

- special procedures for words that contain embedded slashes or hyphens. In the NASA STI Program's system, words containing hyphens are searched once with the hyphen, and, if not found, are then searched without the hyphen;
- recognition of upper and lower case, either of which may be helpful in identifying the meaning of a semantic unit;
- the elimination or addition of broader terms in the same hierarchy as specific terms suggested; and
- the use of a category code to disambiguate terms (such as stress). Ambiguous terms in the text may also be clarified by the choice of entries in the KB. Those semantic units that cannot be disambiguated can generate a new thesaurus term, be left out of the KB, or receive a null translation, that is, be posted to two zeros (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.

Sources of Entries

Sources for the NASA knowledge base 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. Variant forms 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 and phrases that frequently occurred in a Key Word in Context (KWIC) listing of:
  - Titles from documents in the STI database, formerly available on a microfilm cassette as the KWIC Combined File Index.
  - 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. A KWIC listing of phrases that partially matched KB keys, is illustrated in Figure 2. The use of KWIC indexes for identifying repetitious phrases for knowledge base building doubled the production of new entries. When analysts could find no equivalent NASA term(s) for an often-repeated phrase, the KWIC was still helpful in locating deficiencies in the NASA Thesaurus.
- Phrases that were in the Lexical Dictionary of the Defense Technical Information Service (DTIC) and in scope for NASA. (The NASA MAI system is a third generation of the MAI system used by DTIC.)
- 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 and Library of Congress subject headings - selected terms.
- Selected phrases identified by the KBB Text Analysis Tool, described below (Genuardi, 1990). 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 thesaurus terms and to date has been used to analyze about 300 terms in the NASA controlled vocabulary. Most of the input since June 1988 has been identified through the output of the KB Text Analysis Tool.
- Indexer feedback.

Obviously, there are numerous ways of finding potential new knowledge base entries beyond the controlled vocabulary and the official Use references. In the NASA STI Program's experience, the statistical Knowledge Base Text Analysis Tool and indexer feedback are the most efficient.

```
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 2

KBB Text Analysis tool.

The process of identifying the various expressions that are synonymous with thesaurus terms may seem like an infinite task, especially if the list of thesaurus terms is large - and time consumption is not the only drawback to some of the methods tried. The addition of expressions that are essentially either unique or have a very low frequency of occurrence can lead to an unnecessarily large knowledge base.

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 a procedure that NASA developed for identifying words and phrases in natural language that express indexable concepts and that should be translated to NASA Thesaurus terms.

The NASA-developed KBB program is a statistically based text analysis tool. It 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 3 can be described as follows:

- Input text is selected. Generally, this will be the titles and abstracts of a large set of document records (not less than 150 nor more than 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 below 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).

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 3

- 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.

- 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 4. 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. 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. Those terms with both 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. The use of this tool reduced the number of unwanted one-word terms that are suggested by MAI.

Un-edited KBS 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 4

TEXT PROCESSING

Processing of input text by the Access-2 program is done as follows:

- The computer breaks the input text into word strings by stopping at certain punctuation, such as periods, colons, and semicolons, and any predesignated stopword. (See Appendix B for the Stopword List. The selection of stopwords was based (1) on frequency of occurrence in text in NASA's database, (2) on the words' general lack of indexable concepts, and (3) on numerous tests to determine the best tradeoffs.)

- 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 as the first word in a key. 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 one or more Thesaurus terms that will be provided to the indexer as suggested indexing terms.

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

  - 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" (marked with a flag). 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 B) 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 Descriptions and Outcomes:

Mark off 5-word array in title.
Outcome: Only 2 words exist, therefore the array is "Helicopter Noise."

Look up search key "HELICOPTER;NOISE" in KB.
Outcome: Key found. Posting term(s) "AEROACOUSTICS,AERODYNAMIC NOISE,AIRCRAFT NOISE" returned.

No more words exist in the title. Move to the first string in the abstract.
Outcome: The first MAI-selected string 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 string; concatenate word 1 and word 2 to form a search key (in the sample string this would be ACOUSTIC;DATA), and look up search key in the KB.

Outcome: 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).

Outcome: A poisoned (or flagged) 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 string 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).

Outcome: 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").

Outcome: Keys not found.

If an asterisk is found in the posting term field, then 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. Whenever a key is not found and untried words remain in the 5-word array, continue processing combinations that begin with word 1.

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 string and repeat the PROCEDURE described. The remaining arrays for the first string 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"

Outcome: Keys not found.

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

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

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

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

End processing for "BO-105." This was the initial word of a key that successfully matched a key in the KB and provided a NASA Thesaurus term.

Outcome: Three words now remain in the string: "HELICOPTER MAIN ROTOR" and this becomes the new array.

Look up search key "HELICOPTER;MAIN".

Outcome: 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"

Outcome: 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".

Outcome: Key found. Posting term "ROTARY WINGS" returned.

Poison (flag) "ROTOR".

Outcome: The word "ROTOR" may not be used again without an added unpoisoned word or words.

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

Outcome: Key not found.

No more words remain in string.

Outcome: End processing for this string.

Repeat process described for the remaining strings.

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

The second five-word array in the next string, 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".

Outcome: 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".

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

The only other search keys found in the above strings 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".)
Outcome: Key found. Posting term "DESCENT" returned.

Look up search key "BLADE-VORTEX;INTERACTION".
Outcome: Key found. Posting term "BLADE-VORTEX INTERACTION" returned.

Look up search key "CLIMB;999".
Outcome: Key found. Posting term "CLIMBING FLIGHT" returned.

Look up search key "TURBULENT;WAKE".
Outcome: 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 are in the KB, so the hyphens in these words are kept, and the compounds are 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 is dropped, and the compound is treated as two words.

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. As the size of the Knowledge Base has increased with carefully and statistically selected entries, the output has become more acceptable to the indexers.

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 6 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.

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-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.

analytic documents - documents that are cataloged as a single unit (the primary document) and also cataloged as a series of individual documents (subsidiary documents), such as the papers in a volume of conference proceedings.

DTIC - Defense Technical Information Center.

IPS - NASA's 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.

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.

Lexical Dictionary at DTIC - DTIC's master file that accepts input in natural language and provides output in DTIC's Thesaurus terms that express the same concept.

MAI - machine aided indexing.

NASA Lexical Dictionary system - A system for generating NASA Thesaurus terms automatically from any specified input.

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 word or a group of natural language words that express an indexable concept. See also "Search key."

STAR - a monthly publication announcing the Scientific and Technical Aerospace Reports collected and disseminated by NASA on topics that are pertinent to NASA's mission.

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.

REFERENCES

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

Genuardi, M. T. (1990, October). Knowledge-based machine indexing from natural language text: Knowledge base design, development and maintenance. In H. Czap & W. Nedobity (Eds.), TKE'90: Terminology and knowledge engineering, Volume 1. Proceedings Second International Congress on Terminology and Knowledge Engineering, (pp. 345-351). Frankfurt/M., Federal Republic of Germany: Indeks Verlag.

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.

APPENDIX A

HISTORY

DTIC's Role

Paul Klingbiel, first director of NASA's machine aided indexing project, was active for 18 years in linguistic research at the Defense Technical Information Center (DTIC), formerly called the Defense Documentation Center (DDC). While there, he initiated a Lexical Dictionary which became part of DTIC's machine aided indexing system. DTIC's Lexical Dictionary is a data file that translates input in natural language to output in DTIC's Thesaurus terms that express the same concept.

The present DTIC Lexical Dictionary had its origins in the Natural Language Data Base which was established between 1974 and 1979. The core vocabulary of this file was the the DDC thesaurus with the omission of related and hierarchical terms. Natural language phrases with a maximum length of four words were added from machine aided indexing production runs when they did not match an entry already in the Natural Language Data Base (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. 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 Lexical Dictionary 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 Lexical Dictionary was also obtained. When this file was inverted, it provided information on how the NASA Lexical Dictionary system's Knowledge Base (KB), which was founded on the DDC Lexical Dictionary structure, could translate DTIC's Thesaurus terms to those of NASA's Thesaurus. It was helpful, as well, in identifying natural language phrases that could be translated into NASA posting terms. The DDC Lexical Dictionary was built from MAI production output. NASA's 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. Whatever procedure is used, the intent is to build a KB sufficiently comprehensive to translate natural language input to an equivalent output in NASA thesaurus terms in such a manner that the indexers' role is largely editorial.

NASA KWOC and Data Entry

Klingbiel began the KB with a list of NASA Thesaurus 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. However, it was determined later that an alphabetized list of NASA 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. Each authorized posting term and Use reference that appeared in the NASA Thesaurus was coded, given an appropriate logic code, and entered 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 all DTIC terms to NASA terms became operational in June 1983 and was fully described in NASA Contracter Report 3838. During the following year a similar SS project was undertaken for translating to equivalent NASA thesaurus terms the authorized posting terms of the Department of Energy (DOE). This was a much larger task, and while never totally completed, the SS system has been able to translate virtually all of the DOE terms that NASA encounters. The omissions are largely highly specific atomic energy terms and entries for linked DOE terms.

In order to do machine aided indexing of natural language text, NASA first used DTIC's system of identifying indexable concepts by parsing and selecting only noun phrases from text. It did this by assigning syntax to each word in a word string, and applying specific grammar rules. The method is described at length in NASA Contractor Report 4512, 'Machine Aided Indexing from Natural Language Text' under "SYSTEM DESCRIPTION: MAI with Access-1".

This system became operational first for a single file in August 1986, and became available as an online, interactive system for documents without abstracts in October 1988. At this time, documents with abstracts took too long for online use of the system, requiring an average of a minute and a half wait for MAI-suggested terms.

A new method of identifying indexable concepts was designed to eliminate the need for parsing and to shorten processing time. The new program that carried out the identification of these semantic units was called Access-2. Semantic units were identified by proximity and appropriate entries in the KB instead of by parsing and identifying noun phrases. The new method became operational in May 1989 in an overnight batch mode for analytic STAR documents, in March 1990 in a daytime batch mode for other STAR documents, and for all of the main document series in an online, interactive mode in June 1990. The response time has been reduced from 90 seconds to about 6 seconds with the new method of identifying semantic units.

APPENDIX B: 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 C: AUTOMATIC RE-INDEXING OF NASA'S PRE-THESAURUS STI RECORDS

INTRODUCTION

This is a report on the enhancement of pre-1968 NASA STI database records by the addition of automatically generated NASA Thesaurus terms. The report describes the methods used (1) to identify and set aside the uniterms, (2) to provide a translation for each conceptual term, and (3) to generate additional Thesaurus (conceptual) terms through the use of machine aided indexing (MAI) of available text, such as Title fields and Notes of Content.

The NASA Thesaurus was established in 1968. The words and phrases that indexers used prior to 1968 have been compiled into a Subject Authority List (SAL), which contains 27,380 terms. These are of two kinds: (1) conceptual terms, which consist of one or more words that express an indexable concept, and (2) uniterms, which are one-word terms that must be joined with other one-word terms to express an indexable concept. Multi-word terms are generally conceptual. In the pre-1968 indexing, multi-word terms were sometimes used, but they were also routinely broken apart into uniterms.

The construction of the NASA Thesaurus and the implementation of its use in 1968, caused a radical change in NASA's document indexing procedures. The change in indexing necessitated a different technique for searching. In 1989 NASA asked CASI to generate Thesaurus terms for each pre-1968 record in order to simplify searching for the older records (i.e., G-file records) in the NASA STI database. More than 400,000 records were involved. The reindexing of these records with no special or additional resources - which minimized human review - was a significant challenge for machine aided indexing.

Tests indicated that more than 98 percent of the output of the machine aided indexing procedures was appropriate for the documents examined.

EVALUATION

The computer-generated terms for 100 accessions each from International Aerospace Abstracts and Scientific and Technical Aerospace Reports were evaluated. It was determined that:

o 97 percent of the major concepts that were indexed prior to 1968 were successfully translated.

o 98.5 percent of all computer-generated terms were appropriate.

o Of the 1.5 percent terms generated that were not appropriate, .7 percent (or 44 percent of the errors) were due to the fact that the textual data contained in the Imprint and Notes field had no consistent delimiting features.

o 17 percent of the conceptual terms assigned identified valid concepts previously not indexed.

(For additional conclusions, see Phase 4, Task 3.)

Thesaurus terms for G-file records have been put into each record's Major Term field, and the pre-1968 terms are stored in the Data Term field where they continue to be searchable.

PROCEDURES

The NASA Thesaurus term-generation operation for the G-file records was divided into several phases, described below. The initial goal was to identify these phases.

The general approach was to use both the pre-1968 vocabulary terms assigned to G-file records and appropriate textual-field data as input to various processing techniques. The transfer from pre-1968 to current indexing terms was based on translation tables (see Figure C-1) -- the creation of which involved a combination of computational processing and manual analysis. Textual-field data were processed through a modified version of the natural language machine aided indexing (MAI) system.

VOCABULARY ANALYSIS AND CONSTRUCTION OF THE TRANSLATION TABLES

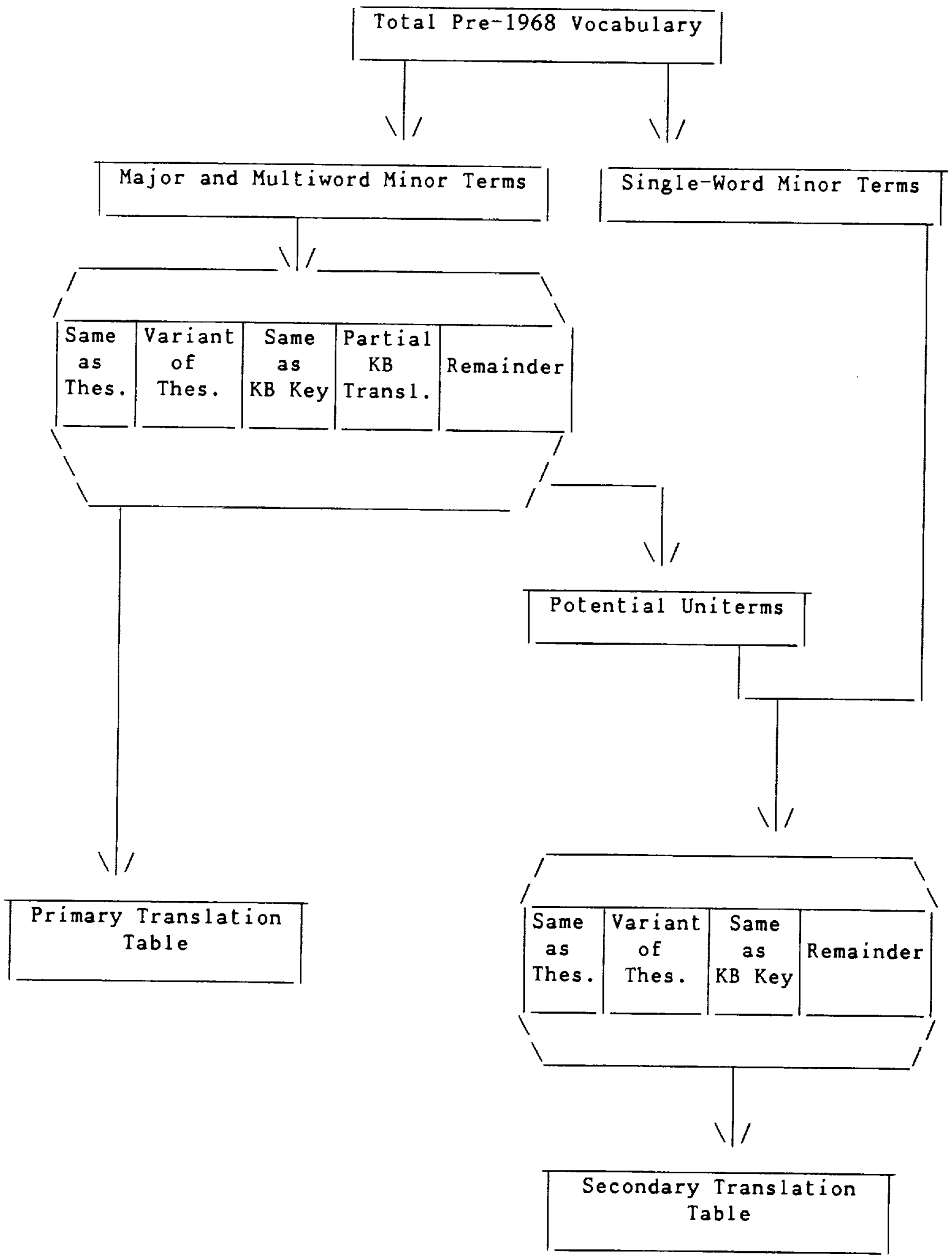


Figure C-1

The target records included all indexed document series currently grouped in the RECON G-file with accession years ranging from 1962 through 1967.

The Major (MJS) and Minor (MNS) Term fields were used for all of the document series with the exception of two files that used only the MNS field. It should be noted that in the series where both fields were used, indexing practice required that the individual words comprising multi-word MJS terms be duplicated as uniterms in the MNS field.

The procedures followed were outlined as follows:

Phase 1 - Analysis and Planning
- Task 1: Analyze G-file vocabulary and record data elements
- Task 2: Develop processing procedures for term translation
- Task 3: Develop procedures for analyzing text

Phase 2 - Construction of the Primary/Conceptual-Term Translation Table
- Task 1: Develop a procedure for the analysis of G-file terms
- Task 2: Apply analysis procedures for the construction of the final pre-1968 term to Thesaurus term translation table
- Task 3: Construct final translation table

Phase 3 - Construction of Secondary Translation Table and Processing the Remaining Single-Word Terms
- Task 1: Apply term analysis procedures for construction of single-word translation table
- Task 2: Construct single-word translation table
- Task 3: Develop procedure for selective translation of single-word terms

Phase 4 - Generation of Pseudo-Records
- Task 1: Define a pseudo-record
- Task 2: Construct the pseudo-record for each accession
- Task 3: Evaluate the results

Phase 5 - Implementation

Phase 1 - Analysis and Planning

The goals of the first phase supported the analysis of the G-file vocabulary, including ways of identifying the differences between uniterms and conceptual terms in the Minor Term field, and devising a general processing plan for generating NASA Thesaurus terms for G-file records.

Task Descriptions

Task 1: Analyze G-file vocabulary and record data elements

Comparative analyses were performed to determine the degree of commonality and translatability that existed between the pre-1968 vocabulary and the current NASA Thesaurus. A series of successive comparisons of the old vocabulary with (1) the current Thesaurus and (2) the MAI Knowledge Base (KB) were carried out.

Major and minor thesaurus term posting frequencies, term lengths, and term word stems were used to identify differences between uniterms and conceptual terms.

A sample of seventy documents was selected for text and indexing analysis. The document selection included samples of all 34 subject categories that existed prior to 1968, documents from each year from 1962 through 1967, and documents from all of the indexed series. For test purposes all titles and notes of content, and some title-like information from other fields were translated to candidate Thesaurus terms using an online MAI routine.

Task 2: Develop processing procedures for term translation

This task defined general processing procedures for term translation and textual data analysis, i.e., per record processing, the pseudo-record format, text processing, and MAI interface. See Figure C-2.

TASK 2 PROCESSES

Procedures for Term Translation

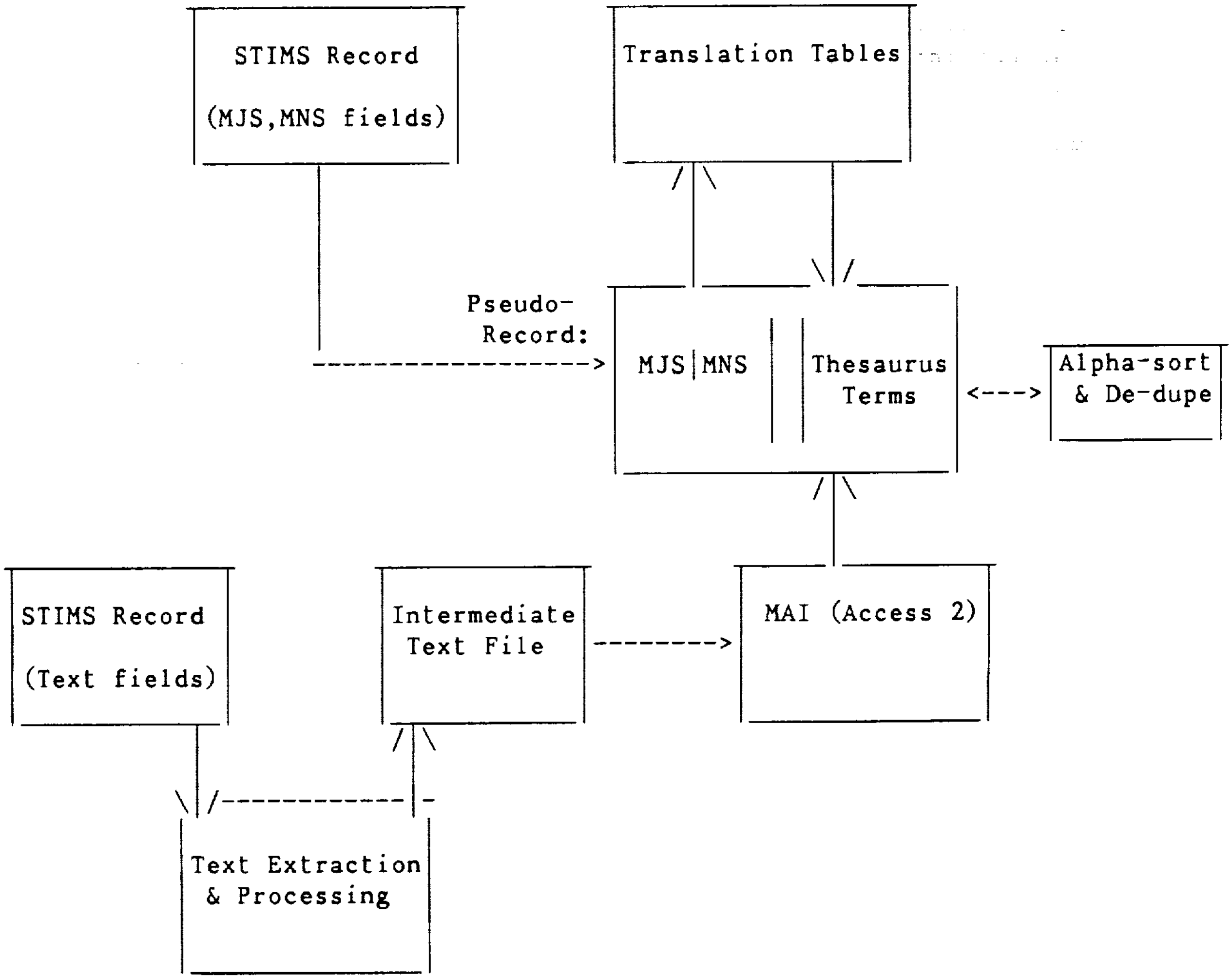


Figure C-2

Task 3: Develop procedures for analyzing text

Task 3 identified fields that contain text to be used for MAI input. These fields were:

* 141	AIN	Imprints and notes (1962-63)
145	UTL	Unclassified title
193	UNC	Unclassified note of content
213	XNC	Textual note of content

* The Imprint and Notes (AIN) field appears in records with accession numbers from 1962 and 1963. The text data of interest is buried within the field in a different position in A-series records and in the N- and X- series records. Thus, a separate extraction algorithm was required for each series group:

A-series Text was extracted starting at the beginning of the field and stopping at the first occurrence of a period followed by a space.

N/X-series The starting point for text extraction was determined by finding the first occurence of a comma then searching ahead 8 words for the last occurence of a "START-WORD." The START-WORD list consisted of 200 cities, states, and countries, and some of their abbreviations that were culled from 1962 and 1963 reports and that represented the most frequently occurring corporate source locations. A list of the start words is provided in Figure C-3. When a START-WORD was found, the extraction began with the first word following the START-WORD. When no START-WORD existed, the extraction began with the second word following the comma (comma+2). Data extraction stopped at the first occurence (after starting point) of either "/U/" or the date year "19xx".

"START-WORDS" - WORDS THAT SIGNALLED THE END OF THE CORPORATE SOURCE CITATION

/AUSTRALIA/
/BELGIUM/
/CANADA/
/ENGLAND/
/FRANCE/
/GT.
/GT. BRIT./
/ITALY/
/N. ZEALAND/
/NETHERLANDS/
/NORWAY/
/ONTARIO/
/PRAGUE/
/SWEDEN/
/SWITZERLAND/
/W. GERMANY/
ALA.
ALBANY
AMES
ANGELES
ANGELES.
ANN ARBOR
ANTONIO.
ARBOR
ARBOR.
ARIZONA
ARIZ.
ARK.
ATHENS
ATLANTA
AUSTIN
BATON ROUGE
BERKELEY
BLACKSBURG
BLOOMINGTON
BOULDER
BOZEMAN
BRIT./
BROOKLYN
BURBANK
BURLINGTON
C.
CALIF.
CALIFORNIA
CAMBRIDGE
CHARLOTTESVILLE
CHATILLON-SOUS-BAGNEUX
CHICAGO
CHICAGO.
CITY.

CLEVELAND
CLEVELAND.
COLLEGE.
COLLEGE PARK
COLLEGE STATION
COLO.
COLUMBIA
COLUMBUS
CONN.
COPENHAGEN
CORVALLIS
D.
D. C.
D.C.
DALLAS
DAVIS
DC.
DEL.
DENVER
DURHAM
EAST LANSING
EL PASO
EUGENE
FAYETTEVILLE
FLA.
FLAGSTAFF
FORT HALSTEAD
FRASCATI
GA.
GAINSVILLE
GOLDEN
GUNNISON
H.
HAIFA
HALSTEAD
I.
IDAHO
ILL.
IND.
IOWA
IOWA.
IOWA CITY
J.
JERUSALEM
KANS.
KINGSTON
KNOXVILLE
KY.
LA HABRA
LA JOLLA

LAWRENCE
LEGON
LEXINGTON
LINCOLN
LIVERMORE
LOGAN
LONDON
LOS ANGELES
MADISON
MAINE
MANHATTAN
MARIETTA
MASS.
MD.
MICH.
MINN.
MINNEAPOLIS
MO.
MONT.
MOTTINGHAM
N.
N. C.
N. DAK.
N. H.
N. J.
N. MEX.
N. Y.
N.C.
N.DAK.
N.H.
N.J.
N.MEX.
N.Y.
NC.
NEV.
NEW YORK
NEWARK
NH.
NJ.
NORMAN
NY.
OHIO
OHIO.
OKLA.
OREG.
OTTAWA
PA.
PARIS
PARIS.
PARK.

PASADENA
PASADENA.
PHILADELPHIA
PHOENIX
R.
R. I.
R.I.
RALEIGH
RAPID CITY
RIVERSIDE
ROCK.
ROME
RUSTON
S.
S. C.
S. DAK.
S.C.
S.DAK.
SALT LAKE CITY
SAN ANTONIO.
SAN DIEGO
SAN FRANCISCO
SEATTLE
ST. PAUL
STATE COLLEGE
STILLWATER
STOCKHOLM
STORRS
TALLAHASSEE
TALLAHASSEE.
TENN.
TEX.
TEXAS
TORONTO
TUCSON
U.
UNIVERSITY PARK
URBANA
UTAH
V.
VA.
VALCARTIER
W.
W. VA.
WASH.
WEST VA.
WIS.
WYO.
Y.
YORK.

Figure C-3

Phase 2 - Construction of the Primary/Conceptual-Term Translation Table

The goal of the second phase was to automate procedures to identify, and separate from the uniterms, any pre-1968 terms that were used for indexing concepts, and to provide each conceptual term with a translation to one or more current NASA Thesaurus terms.

Task Descriptions

Task 1: Develop a procedure for the analysis of G-file terms

Conceptual terms and uniterms are mixed together in the G-file Subject Authority List (SAL), the pre-Thesaurus indexing vocabulary. The general plan was to automatically identify the conceptual terms, group them into a series of subsets that would call for progressively increasing levels of manual review, and that would lead to reasonable, automatic translations to Thesaurus terms. See Figures C-4 and C-5 below.

OVERVIEW OF TERM GROUPING
FOR
THESAURUS TERM GENERATION

Thes. = Thesaurus
KB = Knowledge Base
Key = Knowledge Base field that provides access to equivalent NASA Thesaurus term or terms
Transl.= Translation

The Pre-1968 Vocabulary

Pre-68 Terms, Not One-Word Minor Terms					One-Word Minor Terms			
Same as Thes.	Variant of Thes.	Same as KB Key	Partial KB Transl.	Remainder	Same as Thes.	Variant of Thes.	Same as KB Key	Remainder

Figure C-4

TERM GROUPING AND HANDLING OVERVIEW

PRE-1968 VOCABULARY SUBSET IDENTIFIED	MACHINE ACTION TAKEN	MANUAL REVIEW
TERMS IN EACH SUBSET IDENTIFIED WERE SORTED BY THE NUMBER OF OCCURRENCES		
Single-word terms that occurred only in the Minor Term field	Separated & put in Set A for possible use in the secondary translation table (2nd table)	Yes (Phase 3)
Remaining terms that are the same as Thesaurus terms	Retained in the primary translation table (1st table)	Minimal
Remaining terms same as Thesaurus terms, but with variant endings	Changed to Thesaurus format for the 1st table	Minimal
Remaining terms same as MAI Knowledge Base keys	Translated by MAI for the 1st table	Minimal
Remaining terms that include a KB key(s)	Translated by MAI for the 1st table	Yes
Remaining terms that did not translate	Listed for manual review.	Yes
Single-word terms that MAI posted to zero	Put into Set A for later review for 2nd table	Yes (Phase 3)
Terms identified by human review as being potential uniterms	Put into Set A for later review for 2nd table	Yes (Phase 3)
Set A terms that are same as Thesaurus terms	Retained in the 2nd table	Yes
Set A terms same as Thesaurus terms, but with variant endings	Changed to Thesaurus format for the 2nd table	Yes
Set A terms that are the same as MAI KB keys	Translated by MAI	Minimal
Remaining Set A terms	Listed for manual review	Yes

Figure C-5

Task 2: Apply analysis procedures for the construction of the final pre-1968 term to Thesaurus term translation tables

The following steps were used to break down the population of the pre-1968 vocabulary to differing levels of human review requirements:

- All terms that were listed in the G-file SAL but had no postings were not considered for translation.

- One-word terms that occurred only as minor terms were set aside, in a set designated as Set A, for later analysis.

- The remaining pre-1968 terms were divided into five additional sets. They were created successively such that the members already assigned to one group were not considered for another. The sets were defined in the following way and the terms in each subset were sorted and listed for review in order of their frequency of occurrence:

SET 1 Exact EQUIVALENTS between the pre-1968 terms and the NASA Thesaurus (STT file) terms, including terms that MAI zeroes out, i.e., does not translate.

SET 2 SINGULAR/PLURAL VARIANTS of the pre-1968 terms that matched terms in the Thesaurus, including terms that MAI zeroes out.

SET 3 KB COMPLETE MATCHES (all pre-1968 terms that are translated in their entirety by MAI.)

SET 4 KB PARTIAL MATCHES (all pre-1968 terms that will result in at least one posting term when processed through MAI Access-2; i.e., all terms for which any sub-string of the input term can be translated by MAI.)

SET 5 REMAINDER

The sets, sorted by the total number of postings, were listed in descending order. (Any non-embedded slashes (/) and hyphens (-) occurring in pre-1968 terms were dropped for processing.) Sets 4 and 5 were reviewed for possible new entries to the NASA Lexical Dictionary's (NLD's) Knowledge Base (KB). After the KB was updated, Sets 3 and 4 were recreated. The final, edited versions of Sets 1-5 were compiled into a single file called the primary translation table. It consisted of two fields: (1) the pre-1968 terms, which were listed in alphabetical order, and (2) their Thesaurus equivalents. This file served as the basis for vocabulary switching.

Task 3: Construct final translation table

For review, the pre-1968 terms in each set were printed out with other available information extracted from STIMS fields. Applicable information for each term was presented on a single line and included:

- o total number of postings;
- o total postings as a major term;
- o total postings as a minor term;
- o any symbol that appears in the MAI KB record. These are @ (for an array term), * (for a word or term embedded in a longer MAI KB key), and 00 (for a word or phrase that should not be translated);
- o the (pre-1968) term;

The translations provided by Sets 1, 2, and 3 were generally accepted as accurate, however each set was manually reviewed and edited. The original Sets 3 and 4 were examined for possible new KB entries, and the recreated Sets 3 and 4 were scanned for errors or omissions. Any questionable translation was researched and corrected, if necessary.

Phase 3 - Construction of a Secondary Translation Table and Processing Procedure for the Remaining Single-Word Terms

The goals of phase 3 were to develop a final, automatic process that would discriminate between the usage of single-word terms as uniterms and as conceptual terms, whenever possible, and to translate the latter.

Task Descriptions

Task 1: Apply term analysis procedures for the construction of the single-word translation table

Insofar as possible, the analysis procedures that were used in phase 2 to create 5 sets of processed terms were applied to the single-word terms. (See Figure C-2.)

Task 2: Construct final single-word translation table

As in phase 2, each set was sorted and listed in order of frequency of use. Sets 1 through 4 were manually reviewdd, and, where necessary, corrected. In set 5, the conceptual, single word terms were identified manually. The main criterion for assigning a translation to the pre-1968 single-word terms was whether a given word could potentially have been used as a conceptual index term. This was determined through the analysis of the following factors: the posting frequency total, the ratio of major to minor posting frequencies, and whether the term had a specific meaning when used in isolation. In addition, the KB selection criteria were used as a decision aid in determining which terms were conceptual and which were not.

The pre-1968 terms and their Thesaurus term equivalents that were identified in this phase were assigned to the secondary translation table.

Task 3: Develop a procedure for the selective translation of single word terms

A processing procedure was defined such that the translation of a single-word term would be suppressed if that term was embedded in any of the multi-word NASA Thesaurus terms assigned to a given G-file record.

Phase 4 - Generation of Pseudo-Records

The goal of phase 4 was to define and create for each G-file record a parallel 'pseudo-record' that would contain the output from the automatic indexing process and additional information. These pseudo-records were used to assess the processing procedure and resulting output. After initial assessment of test records, pseudo-records were created for every record in the G-file.

Task Descriptions

Task 1: Define a pseudo-record

A pseudo-record will contain four elements:

(1) a unique G-file accession number;
(2) the MJS terms for that accession, if any (from field 197);
(3) the MNS terms, if any (from field 198); and
(4) a field in which the output from translation-table/ MAI processing will be listed (to be called A-terms). Each A-term list is the per-record output of MAI and translation-table processing.

PSEUDO-RECORD FORMAT

Accession #	MJS	MNS	A terms

Task 2: Construct the pseudo-record for each accession

Each G-file accession number and its associated MJS and MNS terms, which were originally extracted from the STIMS records, were preserved throughout processing. To construct the per-record A-term list, the pre-1968 terms were processed through the primary and secondary translation tables and the designated text was processed through MAI. See example below.

Accession #	MJS terms	MNS terms	A terms
65A10017	Collision parameter	Angle	Collision parameters
	Detonation wave	Collision	Condensing
	Reflected wave	Condensation	Detonation waves
	Wave reflection	Critical	Explosives
		Derivation	Mach number
		Detonation	Mach reflection
		High explosive	Plane waves
		Mach number	Reflected waves
		Parameter	Wave reflection
		Plane	
		Reflection	
		Wave	

Task 3: Evaluate the results

A follow-up statistical analysis was done on 100 1A accessions and 98 1N accessions. 1A accessions are bibliographic records published by the American Institute of Aeronautics and Astronautics in its 'International Aerospace Abstracts' journal; 1N accessions are bibliographic records published by the NASA Center for AeroSpace Information in its 'Scientific and Technical Aerospace Abstracts' journal. The analysis yielded the following conclusions:

- The translation of uniterms was successfully suppressed; i.e., all minor terms that were translated were conceptual and translated to appropriate Thesaurus terms.
- 97 percent of major concepts were translated.
- 98.5 percent of all machine-generated terms are appropriate.
- Conversely, about 1.5 percent of machine-generated terms are errors. Approximately half of these errors are due to the fact that the textual data contained in the Imprint and Notes field had no consistent delimiting features. For 1A documents it was possible to be more accurate than for 1N because the wanted text could be more easily identified.
- The average number of machine-assigned terms per record was nearly 6, which was an average of 1 additional concept per document over those contained in the original indexing.
- 60 percent of the records were enhanced with additional, appropriate, conceptual terms.
- 17 percent of the conceptual terms assigned identified valid concepts previously not indexed.
- Approximately 1 percent of the records had no Thesaurus terms generated. It is reasonable to expect that records without original major terms (8N and 8X) will have a small number of Thesaurus terms, if any, generated by the procedures described above.

Phase 5 - Implementation

Tasks required for implementing the searchability of G-File records included the following:

- For each G-file record, move all major and minor terms into the searchable Data Term field.
- For each G-file record, enter the computer-generated NASA Thesaurus terms into the Major Term field.
- Extend Thesaurus maintenance capabilities to the G-file.
- Advise RECON users of Thesaurus term availability in the Major Term field (197) for pre-1968 records, and the method for continuing to search the G-file with pre-1968 terms which will be stored in the Data Term field (205).
- Contact all recipients of NASA tapes containing STAR, CSTAR, and IAA records explaining the G-file changes. Give each recipient the opportunity to obtain an updated G-file tape.

APPENDIX D

KNOWLEDGE BASE MAINTENANCE

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

- Updates to the NASA Thesaurus (also to the DTIC and DOE thesauri).
- Modifications and additions suggested by NASA indexers.

Updates to the NASA Thesaurus

One problem associated with MAI database maintenance arises from the fact that the conceptual domain, as represented by the controlled vocabulary, is not static. New terms are added regularly and integrated into existing conceptual hierarchies. Each change in the controlled vocabulary requires at least one change to, addition to, or deletion from, the MAI knowledge base. A single change in the thesaurus can cause numerous knowledge base changes. A printed version of the database sorted by the controlled vocabulary terms is a help, but it is unwieldy, time-consuming, hard to read, and quickly becomes outmoded. It became apparent that a new tool was needed to more efficiently identify the areas of the knowledge base that required modification in response to thesaurus changes.

The Knowledge Base Maintenance (KBM) program was created to fill that need. The KBM program is essentially the same as the KBB program except that the final procedure in the KBB process has been altered to allow phrases already translating to a thesaurus term to be included in the output and flagged with two asterisks for easy recognition. Text expressions that have been mapped to an existing thesaurus term in lieu of the newly established term are evident at little more than a glance. The output format of the KBM routine (illustrated below with a few lines of output related to an analysis of the proposed new term MARS CRATERS) includes the ranked phrases with the corresponding thesaurus terms.

```
Occurrences     Phrase                        Term(s)


**   250        CRATERS ON MARS               PLANETARY CRATERS
     108        CERULLI CRATER                --
**   102        MARTIAN IMPACT BASIN(S)       MARS (PLANET),STRUCTURAL BASINS
```

In this case, the KB entries corresponding to the computer-extracted phrases CRATERS ON MARS, and MARTIAN IMPACT BASIN(S) would be updated by replacing the obsolete term translations MARS (PLANET) and PLANETARY CRATERS with the new term MARS CRATERS. CERULLI CRATER, when verified as a crater on Mars, would be added as a new entry to be translated to the new term MARS CRATERS.

Updates to DTIC and DOE thesauri should be researched on RECON for their occurrences in natural language, and, if feasible, added to the KB.

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.

Online maintenance commands are described at length in Appendix E.

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. The chart below indicates the capabilities provided by the maintenance system, along with the command used to carry out each 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.

Maintenance capability:	Maintenance System Command:
Creates authority files	VALSETUP
Validates data file	NASAVAL
Enters update transactions	(See * below)
Loads update transactions	NASALOAD
Prints file, alpha by key	NASAPRNT
Prints file, alpha by postings	NASANVRT
Displays 10 records online	NASAFIND
Provides NLD translations online	NEWACC

*The former update command 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 MAI commands are described more fully in the pages that follow.

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

- 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.

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 checking processes reject entries that have any of the following:

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

An invalid posting term or terms.

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:

```
Word1;Word2$Posting term(s) or 00 or *
Word;00$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:

- o See whether or not the job has run satisfactorily. (The Job Control Language return codes (RC) should all equal zeros.)

- o See whether or not there are any errors listed that must be corrected.

- o Record the number of changes and additions to, and deletions from, the KB. These figures are accumulated for reports.

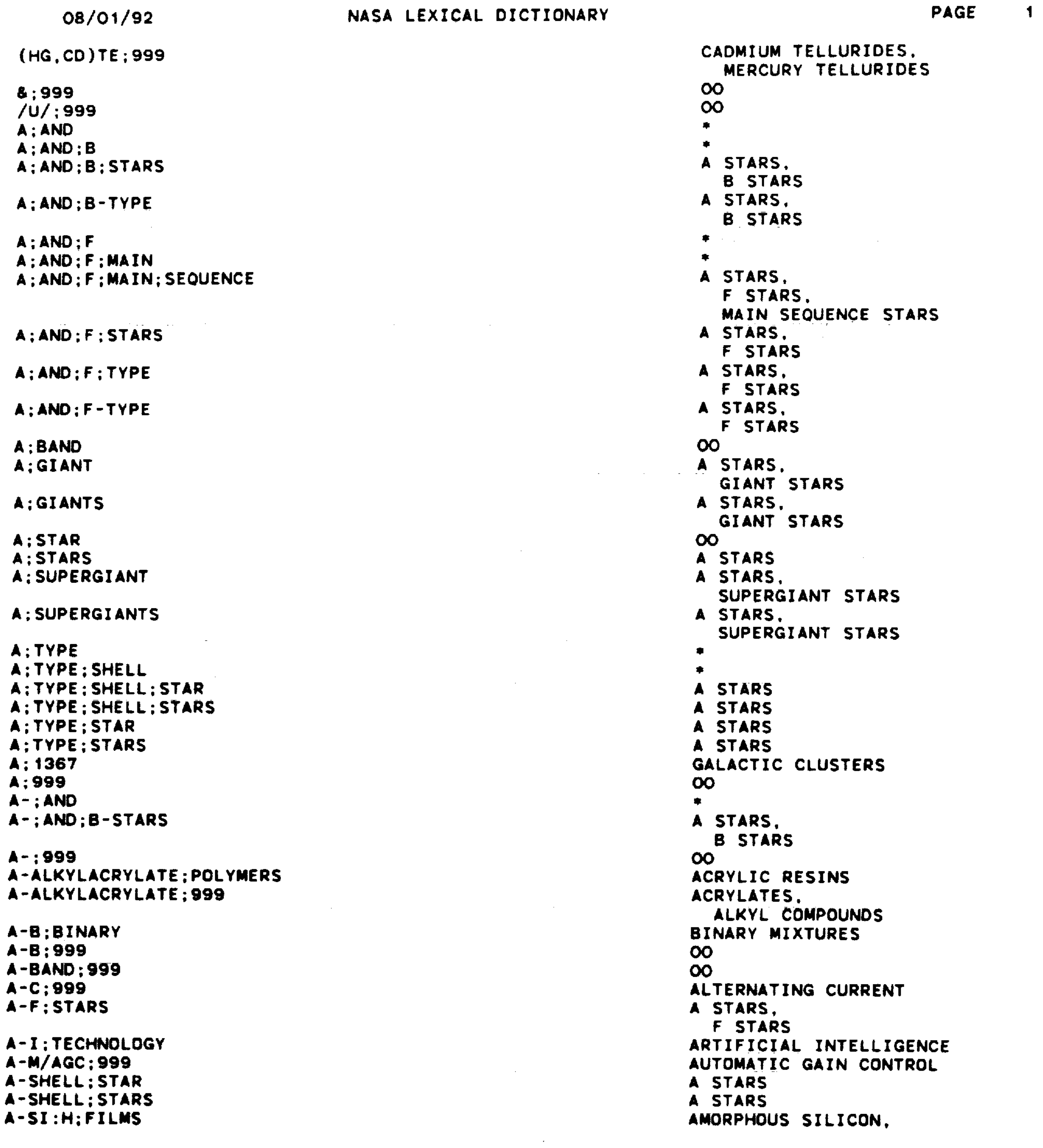
- o See whether or not any KB entry that has been changed needs any further adjustment. For example:

  - 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 leading 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.

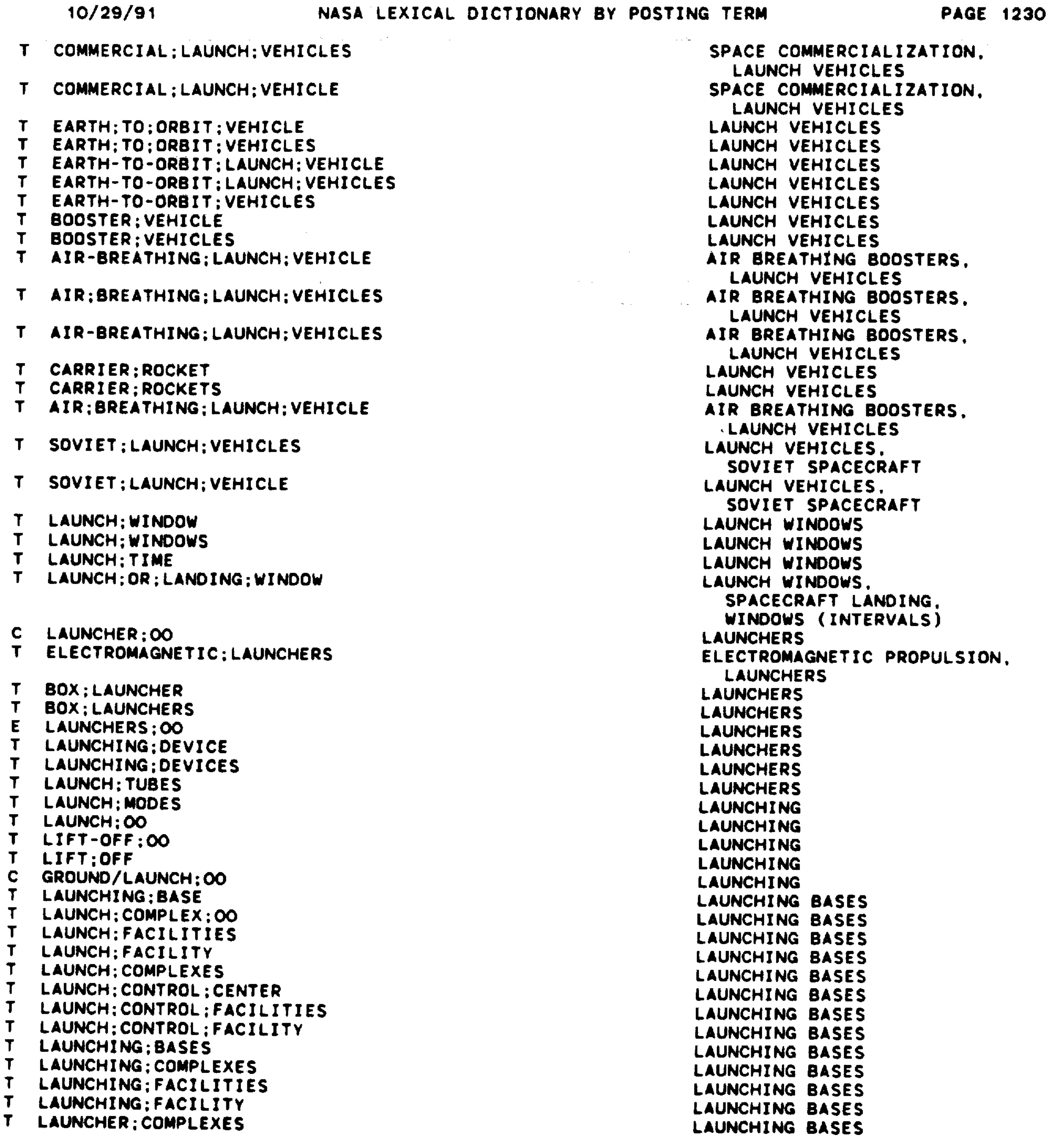
NASANVRT. This command generates a print of the KB sorted alphabetically by posting terms. Before the creation of the KBM program this printout was essential in order to locate a particular posting term in the KB. Entries with multiple posting terms are listed once for each posting term. A sample page 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,
```

NASAPRNT

Figure E-1



```
10/29/91            NASA LEXICAL DICTIONARY BY POSTING TERM            PAGE 1230
 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
```

NASANVRT

Figure E-2

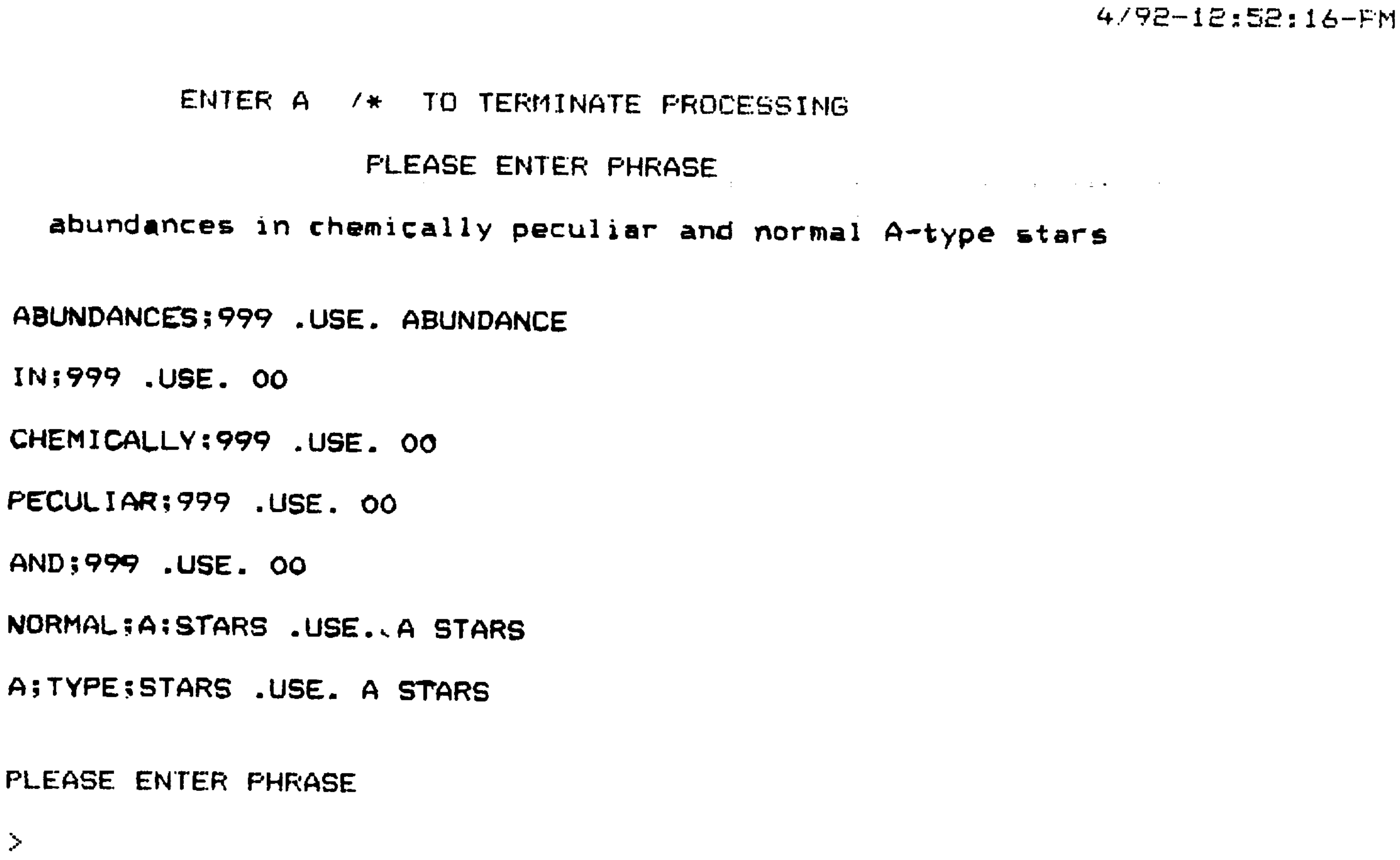
NASAFIND. This command searches the KB for a specified key or first word in a key, and displays ten sequential KB records, beginning with the key or word requested, if it exists. If the requested key or word is not found, the program will locate the sequential position in which it should occur and display 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;of;record')
```

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

 >
```

Sample of NEWACC Output

Figure E-3

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY (leave blank)

2. REPORT DATE
September 1993

3. REPORT TYPE AND DATES COVERED
Contractor Report

4. TITLE AND SUBTITLE
NASA's Online Machine Aided Indexing System
Final Report

5. FUNDING NUMBERS
C NASW-4584

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-4518

11. SUPPLEMENTARY NOTES

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

12b. DISTRIBUTION CODE

13. ABSTRACT (maximum 200 words)

This report describes the NASA Lexical Dictionary, a machine aided indexing system used online at the National Aeronautics and Space Administration's Center for AeroSpace Information (CASI). This system is comprised of a text processor that is based on the computational, non-syntactic analysis of input text, and an extensive 'knowledge base' that serves to recognize and translate text-extracted concepts. The structure and function of the various NLD system components are described in detail. Methods used for the development of the knowledge base are discussed. Particular attention is given to a statistically-based text analysis program that provides the knowledge base developer with a list of concept-specific phrases extracted from large textual corpora. Production and quality benefits resulting from the integration of machine aided indexing at CASI are discussed along with a number of secondary applications of NLD-derived systems including on-line spell checking and machine aided lexicography.

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

15. NUMBER OF PAGES
64

16. PRICE CODE
A04

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 21090-2934
(301) 621-0390