Information Processing & Management Vol. 29, No. 1, pp. 47-59, 1993
Printed in Great Britain.

0306-4573/93 $6.00 + .00

AN OPERATIONAL SYSTEM FOR SUBJECT SWITCHING BETWEEN CONTROLLED VOCABULARIES

JUNE P. SILVESTER
NASA Center for AeroSpace Information, P.O. Box 8757, BWI Airport, MD 21240

and

PAUL H. KLINGBIEL
2435 Sumatran Way, Clearwater, FL 34623

(*Received* 30 *August* 1991; *accepted in final form* 20 *February* 1992)

Abstract – The NASA system of automatically converting sets of terms assigned by Department of Defense indexers to sets of NASA's authorized terms is described. This little-touted system, which has been operating successfully since 1983, matches concepts, rather than words. Subject Switching uses a translation table, known as the Lexical Dictionary, accessed by a program that determines which rules to follow in making the transition from DTIC's to NASA's authorized terms. The authors describe the four phases of development of Subject Switching, changes that have been made, evaluating the system, and benefits. Benefits to NASA include saving indexers' time, the addition of access points for documents indexed, the utilization of other government indexing, and a contribution towards the now-operational NASA, online, interactive, machine aided indexing.

In 1981 the National Aeronautics and Space Administration (NASA) made the decision to begin work on a NASA Lexical Dictionary (NLD). The NLD operates in two modes: Subject Switching (SS) (Silvester, Newton, & Klingbiel, 1984) and natural language Machine Aided Indexing (MAI). Both modes generate candidate NASA Thesaurus terms for the Indexer to review and accept or reject. The NLD System has the following components:

- A now little-used Recognition Dictionary containing syntactic classes of words;
- Three data files that act as translation tables;
- Two programs referred to as access routines (i.e., Access-1, and a second generation version known as Access-2). These manipulate input words and phrases and match them, or subphrases that they construct, with the content of the appropriate data files;
- Applications programs;
- Software used to make the final selection of candidate NASA Thesaurus terms; and
- Programs for NLD maintenance.

This paper describes only the portion of NLD development that applies to the SS of Defense Technical Information Center (DTIC) subject terms to NASA Thesaurus terms (National Aeronautics and Space Administration, 1976, 1982, 1985, 1988, 1991). This portion of SS has been operational since 1983. Its development laid the foundation for the construction of the natural language MAI system in use, online, today; however, while alluded to several times, the online system is not the subject of this paper.

The NLD was developed at NASA's Center for AeroSpace Information (CASI), formerly the NASA Scientific and Technical Information Facility. CASI and DTIC have overlapping interests, and so they share information. About 20% of NASA's database input has already been indexed by DTIC. Documents are received on microfiche accompanied by magnetic tape that provides DTIC's cataloging, abstracting, and indexing in machine-

Requests for reprints should be addressed to NASA CASI, P O. Box 8757, BWI Airport, MD 21240. NASA supports CASI under contract NASw-4584. (No copyright protection is asserted for this article.)

readable form. The NLD was undertaken in order to reduce the reindexing effort that was necessary to adapt the information to the NASA system.

A popular misconception of SS is that it matches the terms of two different vocabularies. This may be the approach used by others, but it misses two critical aspects of NASA's SS system:

1. The NASA SS system matches CONCEPTS not words. As a result, the set of NASA's authorized posting terms selected by the SS system reflects the same concepts that were originally indexed by DTIC with their authorized posting terms. The number of terms in DTIC's set may be totally different from the number of terms in the NASA set, depending upon how many List and Table entries (see the section on Rules, below) are required to effect the switch from one controlled vocabulary to the other.
2. The NASA SS system also takes into consideration the differences between the two agencies' indexing philosophies and practices; for example, DTIC's index set may contain more than one level of a hierarchy of terms, such as Temperature and High temperature. This practice is approved and prescribed by many organizations, but it is not done at CASI. NASA's policy is to index only to the most specific term, in this case: High temperature. Therefore, the SS system is constructed to compensate for such differences.

Paul H. Klingbiel, who had retired from DTIC after 18 years of linguistic research, agreed to organize the NASA project. He had initiated a computer aided indexing project at DTIC that became operational before he left, and had some new ideas for enhancing the system, as well as having a new approach to SS. He began work at the NASA Center in late 1981. Klingbiel's discovery, at DTIC, of the Lexical Dictionary is discussed at length in an article (Klingbiel, 1985, p. 113–126) published in this journal in 1985, entitled "Phrase Structure Rewrite Systems in Information Retrieval." He describes the Lexical Dictionary as a matrix or device constructed to translate input text to authorized posting terms. NASA's Lexical Dictionary (NLD) translates input, in this instance DTIC's posting terms, to NASA's authorized posting terms. The Lexical Dictionary's original discovery was made as a result of a mammoth editing job done on DTIC's Natural Language DataBase. While editing, Klingbiel realized that Use reference construction could be covered with five basic rules, four of which were context free and involved only single words. The context-sensitive rule involved two or more words. These rules are illustrated by the examples below. If the first word of a textual string could be translated without regard to the context, that is, with no additional word or words required to make the meaning clear, then the NLD input had only one word. This was entered into the record in the NLD translation matrix in the Initial Word field. Since additional words were not needed to make the meaning of this word clear, the value of the Final Word field in the matrix was null and was symbolized with two zeros (00).

USE REFERENCE CONSTRUCTION RULES

The Use reference construction rules that Klingbiel applied to the NLD are as follows:

1. Delete: For certain input words or phrases no NASA authorized term is appropriate output. The potential Use reference is therefore null or deleted. This delete rule is symbolized in the NASA system with a zero (0).
   Example:

Rule 0 Initial Word: Fingerprints Final Word: 00
Authorized Term: 00

A shorter way of writing this was soon adopted:
Example:

0 Fingerprints;00 → 00

2. Identity: The input is identical or Equal to the output or the authorized term that should be used. This rule of identity, or an Exact match between input and output, is symbolized with an E.
   Example:

   E Mesons;00 → Mesons

3. Simple Change: A simple Change involves a minor change; the authorized term to be used expresses the same concept, but may, for example, be plural, while the input is in the singular. Or the input and output may have different endings, such as "ing" and "tion," or be synonymous. The symbol for the Change rule is a C.
   Example:

   C Headaches;00 → Headache

4. List: This rule applies when a single concept is translated to multiple authorized terms that, taken together, express the same concept. The symbol for a List is L.
   Example:

   L Machmeters → Mach number,Speed indicators

   Note that when more than one term appears as output, the terms are separated by commas, but no spaces are left, except between words within a multiword term.
5. Table: In the matrix, Tables occur when the input word is context sensitive and requires another word or words to clarify the concept. (Conversely, the Table rule is used when the input Key consists of more than one word.) The second or Final Word in a multiword Table entry may be null. This was originally indicated by 00, because the computers at DTIC sort zeroes last; however, in the NASA system, nines sort last. Since some text contained 99, three nines (999) are used in place of DTIC's two zeroes (00). The object is to have the affected entry end up in last place in the Table when the computer sorts all of the entries that begin with the first word. See Fig. 1 for an example.

We quote from the previously mentioned article by Klingbiel:

> The coding TT represents a Table within a Table. In coding such a Table, provision must be made for phrases consisting of more than two words. A straightforward procedure for accommodating three word and longer phrases is as follows. The first word is placed in the Initial Word column; the second word is placed in the Final Word column. If a phrase is not then complete, an asterisk is placed in the Authorized Terms column to indicate a continuation. The process is continued by combining the Initial and Final Words separated by a semicolon, and the combination is then placed in the Initial Word column (Klingbiel, 1985, p. 118).

Rule	Initial Word	Final Word	Authorized Terms
T	Air	Traffic	*
TT	Air;traffic	Controllers	Air traffic controllers
TT	Air;traffic	999	Air traffic
T	Air	999	Air

Fig. 1. An example of the Use reference construction rule for a Table entry.

Originally there was a progression of symbols used to indicate continuation. A three word entry required a two word entry before it with one asterisk in the Authorized Terms column. A four word entry required two entries before it: one with two words with one asterisk in the Authorized Terms column, and a second with the first two words (separated by a semicolon) in the Initial Word column, the third word in the Final Word column, and two asterisks in the Authorized Terms column. A five word entry used a percent sign to indicate continuation beyond four words. Six and seven word entries used two and three percent signs, respectively, for continuation symbols.

RULES FOR SUBJECT SWITCHING (SS)

In the process of SS from one controlled vocabulary to another, the units in the input are index terms rather than single words. In DTIC to NASA SS, DTIC's Thesaurus terms (Jacobs, 1990) are the input and NASA's Thesaurus terms are the output. The rules remain basically the same as those described above; however, the terms in the Key of a Table entry must be in alphabetical (IBM sort) order, and there is one additional rule.

Rule 1

Delete. For certain DTIC terms there are no conceptually equivalent NASA authorized terms. The potential Use reference is therefore null or deleted and the rule symbol is 0 (zero). For an example, see Fig. 2.

Rule	Initial Term	Final Term	Authorized Terms
0	Fingerprints	999	00
0	Firing rate	999	00
0	Bites and stings	999	00

Fig. 2. An example of a Delete rule.

Rule 2

Identity/Equal. No Final term. Input and output are identical. The rule symbol is an E. For an example, see Fig. 3.

Rule	Initial Term	Final Term	Authorized Terms
E	Mesons	999	Mesons
E	Control sticks	999	Control sticks
E	Acid base equilibrium	999	Acid base equilibrium

Fig. 3. An example of the Identical- or Equal-terms rule.

Rule 3

Simple change. No Final term. Input and output express same concept, are synonymous, or have different endings or different spellings. The rule symbol is a C. For an example, see Fig. 4.

Rule	Initial Term	Final Term	Authorized Terms
C	Intelsat	999	Intelsat satellites
C	Accelerated testing	999	Accelerated life tests
C	Greens functions	999	Green's functions
C	Airport control towers	999	Airport towers

Fig. 4. An example of the Change rule.

Rule	Initial Term	Final Term	Authorized Terms
L	Machmeters	999	Mach numbers,Speed indicators
L	Bituminous coatings	999	Bitumens,Protective coatings
L	Blood plasma substitutes	999	Blood plasma,Substitutes
L	Franz Josef Land	999	Archipelagoes,Arctic Ocean,U.S.S.R.

Fig. 5. An example of a List rule.

Rule 4

List. No Final term. Input is a single index term, while the output consists of two or more terms that might be used to index the same concept. The rule symbol is an L. For an example, see Fig. 5.

Rule 5

Table. Second term required, but it may be null (999) if there are other entries beginning with the same term. More than two terms may be concatenated. An entry joining three terms requires an entry with two terms and an asterisk in the Authorized Terms column. An entry joining four terms must have two entries with continuation symbols, and so on. (Input terms are concatenated in alphabetical order.) For some examples of table entries, see Fig. 6.

Since SS deals with terms rather than with single words, SS Tables combine terms rather than words. In this paper we describe combinations of DTIC terms, but the same principles can be used to translate sets of terms from *any controlled vocabulary* to a set of terms from any other controlled vocabulary, given organizations that index similar categories of material.

As stated earlier, the NASA SS system also takes into consideration the differences in indexing philosophies and practices between DTIC and NASA. For example, when DTIC indexes a document to more than one level of a hierarchy of terms, such as Temperature and High temperature, the NASA system eliminates the broader term. This is done with a Table entry that tells the computer to suggest only the narrower term when both a broad and a narrower term occur in a given set of DTIC index terms. Entries to eliminate broader terms were created largely as the result of Indexer feedback, although they could have been systematically constructed by going through the hierarchical section of the DTIC Thesaurus.

NLD CONSTRUCTION PHASES FOR THE SS FILE

The NLD was constructed by Analysts who were very familiar with NASA index terms, DTIC terms, and both agencies' indexing policies. These Analysts included Project Director Paul Klingbiel and his successors, Senior Indexers, Retrieval Analysts, and the Thesaurus Lexicographer. The Analysts constructed the NLD in four phases.

Rule	Initial Term	Final Term	Authorized Terms
T	Ablation	Nose cones	Ablative nose cones
T	Air	High temperature	*
TT	Air;High temperature	Temperature	High temperature air
TT	Air;High temperature	999	High temperature air
T	Air	999	Air
TT	Coastal regions	Colors	*
TT	Coastal regions;Colors	Scanners	Coastal zone color scanner
TT	Coastal regions	Currents	*
TT	Coastal regions;Currents	Shores	Coastal currents
TT	Coastal regions;Currents	999	Coastal currents
T	Coastal regions	999	Coasts

Fig. 6. Examples of the Table rule.

Input (any terms)	Output (NASA terms)
Chassis	Chassis
Gold	Gold
Gold plate	Cold coatings (a Thesaurus Use ref.)
Gold plated	Gold coatings (an NLD Use reference)
Gold plated chassis	Gold coatings,Chassis
Gold-plated chassis	Chassis (because gold-plated, with a hyphen, had not yet been added to the file or addressed by the program)

Fig. 7. An example of input and output in Mode 1 using the early knowledge base.

Phase 1, the construction of the MAI Knowledge Base (KB), began with entries for every authorized posting term and Use reference in the NASA Thesaurus. When these had been entered, the Analysts added variations of these terms, such as plurals and singulars, synonyms, and other Use references constructed especially for the NLD. Originally, this file was intended for use with both SS and natural language MAI. Due to then-existing storage problems, it was decided, at this point, to construct separate files for these two purposes. (The third NLD file was created later to handle SS of Department of Energy subject terms to NASA terms (Buchan, 1987).)

While the Analysts were constructing the KB, referred to then as the Phrase Matching file, the Programmer was writing a program (Access-1) that could access these files. It was designed to operate in two modes. Mode 1 accessed the KB and attempted to match input, or certain prescribed portions of the input, with entries in the KB. Mode 2 accessed the SS file that was built in Phase 2. An example of input and output in Mode 1, using the early KB, is shown in Fig. 7.

The last item in Fig. 7 (that is, Gold-plated chassis → Chassis) is referred to as a partial match, because only part of the input is translated by the program and the entries in the NLD file. The careful review of all partial matches was an early way of finding needed, new entries for this file. The current method for finding needed entries is based on a statistical examination of text (Genuardi, 1990).

Phase 2, the Subject Switching of Individual DTIC Terms, consisted of the construction of a matrix that would pair each DTIC Thesaurus term with one or more NASA terms that best express the *same concept*. As the Analysts examined each DTIC term, they not only decided whether to use single or multiple NASA terms to express the same concept, but also determined which terms were Not In Scope (NIS) or untranslatable. DTIC terms that had no appropriate translation were considered as having a null posting, expressed as 00 in the Posting Term field. DTIC terms that are Not In Scope for NASA (such as "Area bombing") were posted to NIS. NIS was selected rather than 00, which could have been used, to reassure the Indexers that an equivalent NASA term was unnecessary. For DTIC terms that were translated to 00 (no equivalent concept), it was thought that some appropriate NASA term might surface through feedback from the Indexers to the Analysts. For some examples of DTIC to NASA Subject Switching of individual DTIC terms, see Fig. 8.

Phase 3, the construction of Subject Switching Coordinates, was undertaken after all single DTIC terms were translated. In this Phase, the Analyst looked for groups of terms

Input (DTIC terms)	Output (NASA terms)
Anti fogging agents	Fog dispersal
Antioxidants	Antioxidants
Apogee	Apogees
Approach	Approach+
Architects	Architecture, Personnel
Area bombing	NIS (meaning Not In Scope)
Area coverage	00 (meaning no equivalent concept)

Fig. 8. Examples of DTIC to NASA Subject Switching of individual DTIC terms.

Input (NASA terms)	Output (DTIC terms)
Angular resolution	Angles,Resolution
Gravity wave antennas	Antennas,Gravity waves
Laser interferometry	Interferometry,Laser applications
Intracranial pressure	Internal,Pressure,Skull

Became the following candidate NLD entries:

Rule	Input (DTIC terms; alphabetized and concatenated)	Output (NASA terms)
T	Angles;Resolution	Angular resolution
T	Antennas;Gravity waves	Gravity wave antennas
T	Interferometry;Laser applications	Laser interferometry
T	Internal;Pressure;Skull	Intracranial pressure

Fig. 9. Generating early DTIC to NASA Subject Switching Table entries.

that could be concatenated to form Table entries (i.e., for combinations of DTIC terms that translated to one or more different NASA terms). Since DTIC had a Lexical Dictionary, and was willing to share their files and programs, NASA was able to convert the DTIC system to operate on the NASA mainframe. All of NASA's authorized terms were run through DTIC's Lexical Dictionary. Wherever they used more than one term to express the concept expressed by one of NASA's terms, that entry was reversed, and considered for a DTIC to NASA Table entry. See Fig. 9 for some examples of Subject Switching Table entries.

Phase 4 was concerned with Indexer Feedback and Maintenance. New terms added to either the DTIC or the NASA Thesaurus require additions and modifications to entries in the data files. In addition, users supply feedback as to translations that should be added or modified.

CHANGES

Practical considerations of system operation have a way of changing logically thought-out procedures, and the SS system at CASI is no exception. The first major change was dictated by a change in our computer storage system. It was decreed that each record would have a unique Key that would be the address of the record in its new Virtual Storage Access Method (VSAM) file. This ended sequential searching, ended the distinction between the Initial and Final Word concepts (because to create a unique Key for each record, the programmer combined these two columns or fields into one, separating the units—or elements—with semicolons), and also ended the capability of easily sorting entries by the Final Word field. The change saved storage space and substantially reduced running time.

Another change was made when we realized that the computer needed to read only the first symbol of the logic code when it looked in the file for the logic rule to be followed. Codes of more than one letter, such as Table codes, provide the linguist, but not the computer, with information. To save time, we began coding all kinds of Table entries simply as T.

These changes in operational tactics are very important when manipulating large files run on mainframe computers. The DTIC to NASA SS file today has 18,730 records, occupying 124 tracks, or nearly 6 megabytes of storage. Two DTIC tapes per month are run on appropriate programs in a batch mode on an IBM 4381 mainframe. The Knowledge Base used for machine-aided indexing of natural language text is even larger, with more than 111,000 records, and it is still growing, although the growth rate is now declining. The Knowledge Base currently occupies 581 tracks or nearly 28 megabytes.

A third change occurred when keypunching was eliminated. Originally, when work began on the SS file, Analysts manually printed each entry on coding sheets that were especially designed for the task. Following the coding of entries, Data Entry personnel keypunched cards, sometimes several for a single record. Programmers received the cards, Computer Operators created magnetic tapes from them, and, eventually, the information on tapes was transferred to disks. After some months of this, data entry procedures were updated. Analysts began typing entries directly into the computer. A program verifies input for proper format, and creates a dataset of the transactions. Then, with one more typed command, the computer checks both the Key and Posting Term fields against Thesaurus authority files and automatically loads the records into the master DTIC to NASA SS file. Transactions that contain errors are not loaded, but are written back to the modification dataset, as well as printed out for review. When needed corrections are identified, an Analyst makes them online. Making the first entry directly onto the magnetic tape from a computer terminal obviously saved, and continues to save, considerable time.

FLAGS

As the Analysts examined each DTIC term and decided upon the NASA equivalents, they also flagged the suggested NASA authorized posting terms to provide additional information to the Indexer.

1. At sign (@):
   Example:

   E Alkali metal compounds;999 → Alkali metal compounds@

   NASA Indexers are encouraged to avoid certain broad terms that may be ambiguous or lack specificity and to use a narrower term (NT) instead. These terms-to-be-avoided, known as array terms, are flagged with an "at sign" (@), which follows the pertinent NASA term. If the Indexer consults the NASA Thesaurus, the following scope note will be found:

   (Use of a more specific term is recommended – consult the terms below).

   The non-array NTs listed in this case are: Cesium compounds, lithium compounds, potassium compounds, rubidium compounds, and sodium compounds. If none of these is appropriate, then the Indexer may use the array term (but without the @).
2. Greater than (>) sign:
   Example:

   C Alligator weed;999 → Plants (botany)>

   If the DTIC term has no equivalent NASA term and is translated to a broader NASA term, that NASA term is flagged with a "greater than" (>) sign.
3. Plus (+) sign:
   Example:

   E Bolts;999 → Bolts+

   If NASA has a broad term that is the equivalent of a given DTIC term, and also has narrower terms than the one used by DTIC, and DTIC has no equivalent NTs, then the broad DTIC term is translated to the equivalent broad NASA term and flagged with a "plus sign" (+). This indicates that NASA has narrower terms which should be considered. The Indexer must determine whether or not an available, narrower term is more appropriate than the broad term supplied by the system. A check of Bolts in the NASA Thesaurus reveals that NASA has two narrower terms: Rockbolts, and tie bolts. If either of these is a better choice than the broader term Bolts, then the narrower term should be used.
4. Question mark (?):
   See Fig. 10 for examples.

I Performance tests;999→Performance tests?
I Surface truth;999→Ground truth?,Sea truth?
I Space surveillance systems;999→Space surveillance (spaceborne)?,
Space surveillance (ground based)?

Fig. 10. Some examples of Indeterminate entries.

When the Analyst cannot tell the Indexer precisely which NASA term to use, but can present one or more possibilities, a question mark (?) follows each NASA posting term suggested as a translation for the input DTIC term. The Rule or Logic Code for this kind of entry is an I, which indicates that this translation is context sensitive and therefore Indeterminate; the choice must be made by the Indexer based on the document at hand. In the first example, the Analyst cannot tell if this match of characters constitutes a match of concepts or not. The NASA term applies only to operating equipment. DTIC uses the term more broadly, covering not only performance of operating equipment, but also performance of humans and animals. The Indexer must determine whether the NASA terms applies to the document at hand or not. In the second and third examples, NASA has no equivalent term, but does have two, more specific terms. The Indexer must determine which narrower term is correct, based on the content of the document.

FINDING A TRANSLATION

All DTIC terms have some kind of translation in the NLD, even if null because no translations exist, are wanted, or are in scope. When the computer looks in the NLD for an input DTIC term, the pointer goes to the spot just ahead of that term's first occurrence in the Initial Term position in a Key. The computer then reads the Rule symbol (referred to at NASA as the Logic Code) for the first entry and follows the Rule indicated to find a NASA translation. Take, for example, the DTIC term "Frequency." The computer will find a T in the Logic Code or Rule field, which tells the computer to look for another term to follow "Frequency." The entire NLD Table of entries beginning with this DTIC term is shown in Fig. 11.

If the document's set of posting terms does not include any of the DTIC terms listed in the Final Term position, then the program defaults to the entry that has 999 as the Final Term and provides the NASA term "Frequencies" as output.

Table entries in the database are entered with the Initial Terms and Final Terms in each Key in alphabetical order. Before the computer tries to identify Table entries in a set of index terms, the index terms that constitute the set are sorted alphabetically. Only then are the terms for that document combined and tried against the NLD entries. Since the concatenated terms are also in alphabetical order, they will match any Key containing those terms.

Rule	Initial Term;Final Term (=Key)	Authorized Postings
T	Frequency;Infrasonic radiation	Infrasonic frequencies
T	Frequency;Ionization	Ionization frequencies
T	Frequency;Plasmas physics	Plasma frequencies
T	Frequency;Scanning	Frequency scanning
T	Frequency;Stability	Frequency stability
T	Frequency;Synchronization electronics	Frequency synchronization
T	Frequency;Vibration	Vibrational spectra
T	Frequency;999	Frequencies

Fig. 11. DTIC to NASA Table entries beginning with the DTIC term "frequencies."

Besides alphabetizing terms before they are examined, the computer does another "housekeeping" chore to prepare the input for comparisons. Any term that contains a gloss has the parentheses removed. DTIC's glossed terms have no space between the word and the left parenthesis of the gloss. NASA's terms do have a space. By dropping the parentheses and creating the Key: Stress physiology;999 the computer determines that DTIC's: Stress(physiology) now written: Stress physiology;999 should be translated to NASA's: Stress (physiology)+. Of course, because of the plus sign, the Indexer may select some narrower term.

BENEFITS

Benefits obtained from the use of the NLD were measured with the least possible disruption to the indexing process. The hypothesis upon which the NLD was authorized was that the NLD would increase the Indexers' productivity and reuse the indexing already done by DTIC. It was intended that the quality of the indexing would remain high. The following analyses were used to test the adequacy of our sample, the significance of our results, and the proof of our hypothesis.

Evaluation methods

The evaluation of the NLD was based on a comparison of the preliminary subject analysis study done for the period December 1982 through March 1983 with a post implementation study done for the period December 1983 through March 1984.

Study number 1 included confidential interviews with each Indexer; all were conducted by the same interviewer to ensure consistency. In addition, a sample of 100 documents was selected from a single DTIC magnetic tape from which NASA selects input for the NASA database. Such tapes are received twice a month. The results of SS for these 100 documents were analyzed. The 100 sample documents were taken from all subject categories represented on the tape. Since there were fewer than 100 categories, multiple selections were made from some categories in approximate proportion to the number of documents assigned to the more populous categories.

Study number 2 utilized a questionnaire because there was concern that observed time studies would be intrusive and slow production. Indexers, without consulting with one another, filled out their questionnaires simultaneously. In addition, a representative sample of 150 DTIC documents drawn over a three-month period was analyzed.

Comparisons

See Fig. 12. Although study 1 had a sample of 100 documents, two of the DTIC posting term values were discarded as being too deviant, leaving a sample size of 98.

Figure 13 shows some of the comparisons that were made between DTIC's indexing and NASA's manual and Subject Switched indexing.

The standard error is small. Since the *t* test (used to ascertain the deviation of the estimated mean from the mean of the population) gives us a value off the *t* chart that indicates a better than 99% confidence level, we conclude that our samples and the results of our comparative study (shown in Fig. 13) are valid for the entire population.

It is interesting to note that before using the NLD, there was considerable difference in the average number of index terms assigned by the two agencies: 14.32 to 9.59. Study 2 shows that the averages are now very close: 13.09 to 12.60.

Access points, productivity, and retrieval

The increase in the number of NASA index terms assigned to a document, as indicated in Fig. 13, not only signals increased productivity, but also increases the number of access points to a document. Evaluation forms filled out by recipients of searches indicated that these indexing changes did not affect retrieval adversely, and that the pertinency level of retrieval remained high throughout the introduction and use of the NLD.

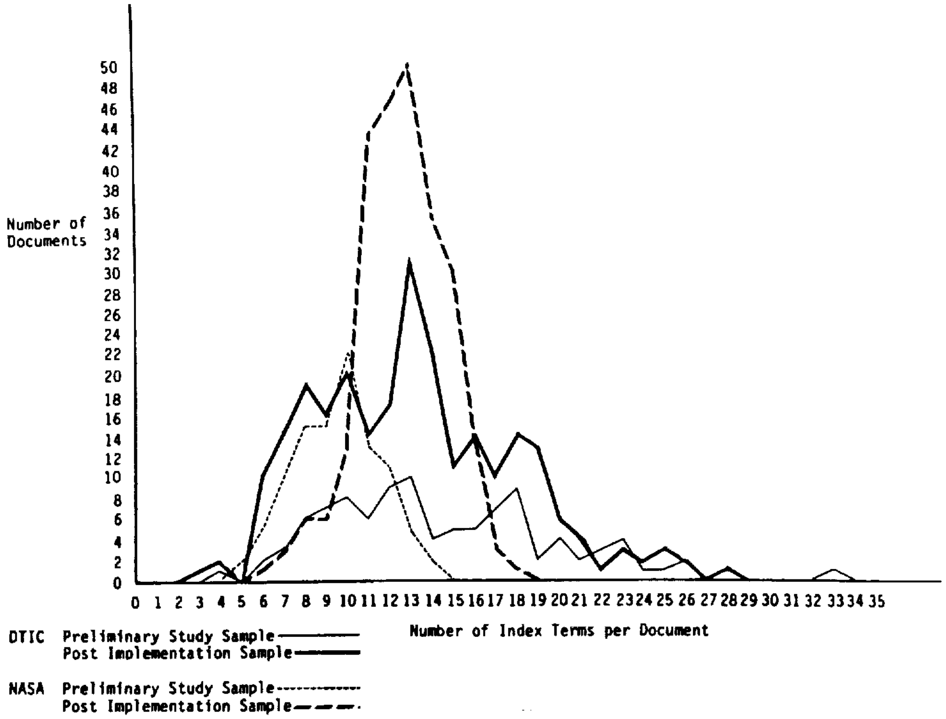

Fig. 12. A comparison of DTIC and NASA indexing.

Time savings

Two concerns in the field of library and information science are the ever-growing amounts of material to be classified, stored, and disseminated, and a constant need to do more work for less or the same amount of money. Information scientists are looking for ways to get information to the user more quickly. We feel that the NLD is making a contribution in this area.

Seventy percent of the Indexers reported having index terms provided by the NLD makes indexing DTIC documents faster. The remaining Indexers indicated that having the suggested NLD terms has no effect on their speed. Indexers were asked to estimate the time saved by having NLD terms. The average of the estimated savings was 5.4 min per document (see Fig. 14).

		Study 1 (Pre)		Study 2 (Post)	
		DTIC	NASA	DTIC	NASA
Documents in sample	N	98	100	250	250
Mean of term assigned	$\bar{X}$	14.32	9.59	13.09	12.60
Standard deviation where Σ = the sum and X = the deviation from the mean	$S = \sqrt{\frac{\Sigma x^2}{N-1}}$	4.88	2.03	4.73	2.05
Variance	$S = \sqrt{\frac{x^2}{N-1}}$	23.81	4.11	22.35	4.19
Standard error of mean	$S_{\bar{x}} = \frac{S}{\sqrt{N}}$	.49	.20	.30	.13

Fig. 13. Comparison of DTIC's indexing with NASA's manual and Subject Switched indexing.

Estimate of time saved

Including the value of zero for each Indexer who did not save time, we have the values of 3, 2, 10, 4, 3, 1, 15, 0, 0, and 0 min saved per document. The average (total min/10) saving is 3.8 min.

For those who reported a savings, the average is 5.4 min per document.

Estimates without the outliers (i.e., the extremes): 13/5 or 2.6 min saving per document.

Pooled estimate of time saved: 5.4 + 2.6 min/2 or 4 min saving per document.

Estimate of time to index (10 Indexers)

Individual replies were 6, 8, 7.5, 20, 4, 15, 7, 10, 12.5, and 10 min per document for an average time to index of 10 min per document.

Fig. 14. Indexers' estimates of time savings with Subject Switching.

Indexers then were asked to estimate the time required to index a DTIC document with NLD terms provided. The average of these estimates was 10 min. When this information was compared with Study 1 (pre-NLD) average indexing time of 13 min, a 3-min difference was noted. The predicted savings per document was 2 to 3 min. Based on the indexers' estimates, the intended goal has been reached and may have been exceeded. This time savings obviously speeds up the document turnaround time and thus increases the timeliness of the product.

Changes in work emphasis

As an Indexer tool, the NLD has relieved the Indexers of having to look up many terms in the thesaurus. The correct form is presented for use or for deletion. The largely mechanical lookup part of indexing has been replaced with a more intellectual task of watching for coordinations of DTIC terms that can be added to the NLD. The process results naturally from the Indexers' review of index terms presented by the NLD printout; however, this change has provided additional challenge to the Indexer's intellect.

Shared resources

It is wasteful of government resources to reindex documents already satisfactorily indexed at taxpayers' expense. The original and primary purpose of the NLD was to utilize indexing done by other agencies. The sharing of indexing with DTIC also brought about sharing of programming, and ultimately improved quality in the thesauri and lexical dictionaries of both DTIC and NASA.

Stepping stone

The Lexical Dictionary has been a stepping stone to other endeavors. Its Knowledge Base has been and is continuing to be expanded for new applications such as a spinoff spelling check, the addition of NASA Thesaurus terms to MARC records or to pre-Thesaurus-indexed records, and the identification of needed Thesaurus additions. NLD research most notably has supported the development of machine-aided indexing (MAI) based on the analysis of natural language text in titles and abstracts. This system is currently in operation at CASI in an online, interactive mode.

CONCLUSIONS

In the NLD, the NASA Center for AeroSpace Information (CASI) has a system that translates words and phrases from input material into equivalent concepts expressed in NASA posting terms. The system was designed particularly to allow the reuse of DTIC indexing in the NASA environment, which it has done well since 1983. When Subject Switching is abandoned, it will be because NASA's natural language MAI has become more efficient than Subject Switching.

REFERENCES

Buchan, R.L. (1987, Summer). Computer aided indexing at NASA. *Reference Librarian No. 18: Current trends in information: Research and theory* (pp. 274-276).

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. 339-344). Frankfurt: Indeks Verlag.

Jacobs, C.R. (Comp./Ed.). (1990, September). *Defense Technical Information Center thesaurus.* (DTIC Report No. AD-A226000). Alexandria, VA: DTIC, Defense Logistics Agency.

Klingbiel, P.H. (1985). Phrase structure rewrite systems in information retrieval. *Information Processing & Management, 21*(2), 113-126.

National Aeronautics and Space Administration. (1991). *NASA thesaurus supplement.* (NASA Special Publication No. 7064-Suppl-5) Washington, DC: NASA. (NTIS No. N91-19962).

National Aeronautics and Space Administration. (1988). *NASA thesaurus* (*Vols. 1-3*). (NASA Special Publication No. 7064). Washington, DC: NASA. (NTIS Nos. N89-13302; N89-13298; and N89-13301).

National Aeronautics and Space Administration. (1985). *NASA thesaurus* (*Vols. 1-2*). (NASA Special Publication No. 7053). Washington, DC: NASA. (NTIS Nos. N86-20168; N86-20169).

National Aeronautics and Space Administration. (1982). *NASA thesaurus* (*Vols. 1-2*). (NASA Special Publication No. 7051). Washington, DC: NASA. (NTIS Nos. N83-10980; N83-10981).

National Aeronautics and Space Administration. (1976). *NASA thesaurus* (*Vols. 1-2*). (NASA Special Publication No. 7050). Washington, DC: NASA. (NTIS Nos. N76-17992; N76-17993).

Silvester, J.P., Newton, R., & Klingbiel, P.H. (1984, October). *An operational system for subject switching between controlled vocabularies: A computational linguistics approach* (NASA Contractor Report No. 3838). Washington, DC: National Aeronautics and Space Administration. (NTIS No. N85-11903).