AUTOMATING METADATA TAGGING: COMBINING MAN/MACHINE INPUT FOR OPTIMAL OUTPUT

Katrina E. Ricci, John C. Hodak, and Sandra Hughes
NAVAIR Orlando Training Systems Division
Orlando, FL
katrina.ricci@navy.mil; john.hodak@navy.mil; sandra.hughes@navy.mil

ABSTRACT

There is an ever-increasing need to apply metadata to legacy electronic training material as well as to content currently under development. Metadata, or very simply data about data, provide an underlying description of training material. Metadata describe attributes of learning objects including, but certainly not limited to, the content itself, when it was created, who created it, and its intended purpose. This information can allow developers to search and find previously developed content in order to achieve a financial efficiency through updating or reusing existing content. Further, as the future vision of Navy training matures, metadata can help ensure that sailors receive the right training at the right time based on knowledge of an individual sailor's needs and applicable training material.

Metadata are comprised of both objective and subjective data elements. Objective elements are those that are relatively straightforward to identify. They include data such as the developer, the training title, or the revision number of the content. Subjective elements – arguably the more valuable data – more thoroughly describe the training content. However, they are subject to individual interpretation and present a potential time consuming and expensive component to generating metadata. It is very appealing, therefore, to apply automation to the process of generating metadata. Technologies are available to assist in this process. Most notably, the application of a machine learning technology, Latent Semantic Analysis (LSA), can assist in the very arduous task of identifying subjective metadata tags.

This paper will describe the use of LSA in automating the metadata tagging process. Further, results of a research effort examining the use of LSA for metadata tagging will be presented. The results of this study indicate that the most efficient and effective process of tagging electronic training content may be to allocate that function between both the human and the computer.

AUTHORS

Dr. Katrina E. Ricci is a Senior Research Psychologist with NAVAIR Orlando Training Systems Division. She received her M.S. in Industrial/Organizational Psychology and Ph.D. in Human Factors Psychology from the University of Central Florida. Dr. Ricci is the Principle Investigator for the Office of Naval Research, Capable Manpower, Future Naval Capability Advanced Technologies for IETM Development and Delivery program. Her research interests include performance support and training technologies.

Mr. John C. Hodak is a Research Psychologist with NAVAIR Orlando Training Systems Division. He received his M.S. in Organizational Management from the University of Central Florida. He has over 10 years of organizational and managerial experience. Mr. Hodak has worked as investigator on projects involving Interactive Electronic Technical Manuals and multimedia learning.

Ms. Sandra Hughes is a Research Psychologist with NAVAIR Orlando Training Systems Division. She has conducted and/or managed research and development projects in the areas of team performance; stress and decision-making; and distance learning design and implementation. She received a Master's degree in Industrial/Organizational Psychology from the University of Central Florida in 1990.

AUTOMATING METADATA TAGGING: COMBINING MAN/MACHINE INPUT FOR OPTIMAL OUTPUT

Katrina E. Ricci, John C. Hodak, and Sandra Hughes
NAVAIR Orlando Training Systems Division
Orlando, FL
katrina.ricci@navy.mil; john.hodak@navy.mil; sandra.hughes@navy.mil

BACKGROUND

The Navy's Integrated Learning Environment (ILE) is planning and developing an ambitious effort to dramatically change the structure, assembly and delivery of learning content for the Navy. These changes are the result of efforts over the last several years examining the structure and function of the Navy's training strategy. In particular, the Executive Review of Navy Training (ERNT) derived three guiding principles to support an efficient and agile learning environment. These principles include the development of a systematic approach to education and training based on the science of learning; a continuum of learning throughout a sailor's career; and continuous matching of education, training, and job assignments to the skills needed for career and personal development (ERNT, 2001). As a key enabler of these principles, ILE will provide the framework and processes that will improve individual and team performance directly linked to mission essential tasks by making knowledge available to sailors and the fleet when and where it is needed (NPDC, 2004a).

To meet the above goals, the learning content within ILE will be structured to the Navy's Sharable Content Object Reference Model (Navy-SCORM). While SCORM 2004 is a collection of specifications adapted from multiple sources to provide a comprehensive suite of e-learning capabilities that enable interoperability, accessibility and reuse of Web-based learning content, Navy-SCORM relies on the extensible nature of SCORM 2004 to adapt to the specific needs of Navy training. For the Navy, this means the application of the Navy's Content Object Model (NCOM) (see Figure 1).

The NCOM gives meaning to and describes the relationships between assets, enabling learning objects (ELOs), and terminal learning objects (TLOs) within the hierarchy (NPDC, 2004b). More specifically defined, an asset is a single media element or a single text element. An ELO is an aggregation of one or more assets. A TLO is an aggregation of one or more ELOs.

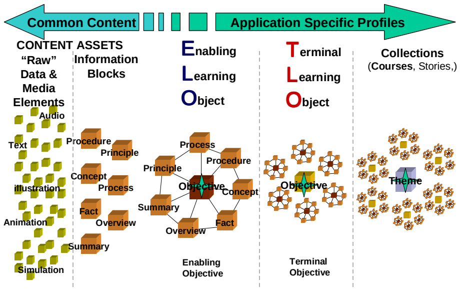


Figure 1: The Navy's Content Object Model

The primary payoff for Navy-SCORM is the ability to reuse and repurpose the wealth of training assets. However, in order to reuse these assets, they must first be found. Repositories of training content may contain huge amounts of data - thus, the ability to find previous assets is essential to achieving potentially huge time and cost savings. Metadata provide the content descriptors that assist in this task.

Metadata

With the future vision of ILE in mind, there is an ever-increasing need to apply metadata to emerging and legacy electronic training material. Metadata, or very simply, data about data, provide an underlying description of training material. Metadata describe attributes of learning objects including, but certainly not limited to, the content itself, when it was created, who created it, and its intended purpose. This information can allow developers to search and find previously developed content in order to achieve a financial efficiency through updating, reusing or repurposing that content. Further, as the future vision of Navy training matures, metadata can help ensure that sailors receive the right training at the right time based on knowledge of an individual sailor's needs and available, applicable training material.

Metadata are comprised of both objective and subjective data elements. Objective elements are those that are relatively straightforward to identify. They include data such as the developer, the training title, or the revision number of the content. Subjective elements – arguably the more valuable data – more thoroughly describe the training content. They include data such as keywords and summary information that details what the content is about. However, generating subjective elements is a much more time consuming and thus, expensive, process. Moreover, subjective elements are subject to individual interpretation – characterized as the "Vocabulary Problem."

The Vocabulary Problem

A great amount of research has addressed the issue known as the vocabulary problem – the problem that two people use the same term to describe a text object less than one-fifth of the time (Furnas, Landauer, Gomez, and Dumais, 1987). In a study published almost 20 years ago, Furnas et al. examined word choice for objects in five domains (text editing operations, message decoding, common objects, classified ads, and recipe keywords). In each domain, participants were asked to supply descriptors for the functions or objects represented in the domain with the expressed direction that the goal was to generate descriptors that would be helpful to other people who would later retrieve or find these objects. Results showed that in every case, people favored the same term with a probability under .20.

Even prior to the work by Furnas et al., many studies have found poor agreement in the assignment of indexing terms, even when subject matter experts are used to generate the terms (e.g., Cooper, 1969; Tarr & Borko, 1974; Tinker, 1966). As Gomez, Lochbaum, and Landauer explain, different people – or the same person on different occasions – will be interested in different aspects of the same object. Thus using subject matter experts to find the "right" names is impossible as a single name or small set of names will fail to serve many retrieval purposes. Rather, Gomez et al. demonstrated that search success is improved by greatly increasing the number of names per object.

It is very appealing, therefore, to apply automation to the process of generating the huge amount of metadata needed. The application of a machine learning technology, Latent Semantic Analysis (LSA), can assist in the very arduous task of identifying subjective metadata tags.

Latent Semantic Analysis (LSA)

LSA is a statistical technique for describing and comparing the similarity of bodies of text. It does this by applying an automatic technique for extracting and inferring relations of expected contextual usage of words in passages of text. It does not require manually constructed dictionaries or ontologies. Rather, it uses only raw text that is parsed into words, and separated into meaningful passages such as sentences or paragraphs.

LSA is based on a mathematical technique closely akin to factor analysis. First, LSA uses the frequency with which two words are used within a portion of a text to establish a probabilistic measure of semantic association. In the first step, the text is represented as a two-dimensional matrix in which the rows stand for unique words and each column stands for another word, a sentence, a paragraph, or some other context. Each cell contains the frequency with which the word of its row appears in the passage denoted by its column. Thus, the matrix has the form of a large rectangular table with a large number of rows and an even larger number of columns. The cell entries are then weighted by a function that expresses the word's importance in the passage and the degree to which the word type carries information in the domain.

Next, LSA applies a statistical technique called Singular Value Decomposition (SVD) that compresses the co-occurrence information into a smaller space. In SVD, a rectangular matrix is decomposed into the product of three other matrices. The first derived matrix describes the original row entities as vectors of derived orthogonal factor values. The second derived matrix describes the original column entities in the same way. The third matrix is a diagonal matrix containing scaling values such that, when the three component matrices are multiplied, the original matrix is reconstructed. The final step is to identify relationships that underlie the pattern of occurrence of words across passages. In theory, semantically related words and passages will load on similar dimensions, although they may share no common words.

The idea underlying this approach is that the similarity of the contexts in which a word appears is a reliable indicator of the similarity of the meaning of words to each other (Landauer, Foltz, Derr & Leacock, 2006). By processing a large sample of language, and specifying the contexts in which words occur, LSA can organize any set of these words in a high dimensional semantic space. Empirical evidence shows that LSA estimates of the similarity of text meanings overlap about 90% with human domain expert judgments (Landauer 2002).

The technological basis of the prototype metadata tagging system used for the current research is a combination of machine-learning technologies, such as LSA, and a user interface to facilitate confirmation and editing of automatically generated content tags and classifications.

The raw text used to "teach" the system was a large body of general English text (roughly 14 million paragraphs) as well as domain specific text (17,000 Naval engineering curriculum documents).

The prototype software supports the metadata tagging by generating the following:

Best sentences: The meanings of the sentences of a sharable content object (SCO) are compared with each other and that of the whole, and an algorithm chooses k (the number of sentences is selectable by the user from a pull-down menu) sentences such that the overall meaning of their combination most nearly matches that of the whole.

Summary keywords: The summary keywords algorithm finds n (the exact number is selectable by the user from a pulldown menu) words chosen from the training corpus (not just the SCO) that together best approximate the SCO's total meaning.

Categorical classification: The system is trained on examples of SCOs previously assigned by experts to categories in a predetermined, e.g. hierarchical, classification scheme. New SCOs are assigned 0-100% similarities with every category. Two taxonomies were used: the Defense Technical Information Center (DTIC) Subject Categorization Guide and the Standard Subject Identification Codes (SSIC).

The current study addressed the ease of use of the demonstration tool and the overlap of keywords generated by study participants and the output of the tool.

METHOD

Participants

Ten (10) participants, all employees or interns in the Human Systems Research and Engineering Division, Human Performance Research and Integration Branch of NAVAIR Orlando, took part in this investigation. Seven (7) participants were male and three (3) were female. All participants had at least a Bachelors degree, five (5) participants had some post-Bachelor's study, two (2) had Master's degrees, one (1) had conducted post-Master's study, and one (1) had received a Ph.D.

Materials

Three excerpts from the Navy's Combat 'A' curriculum were used as sample texts for this study. The excerpts covered basic material from the areas of security, corrosion control, and hydraulics. Each sample was approximately one paragraph long (100 to 150 words). As an example, the security text is provided below:

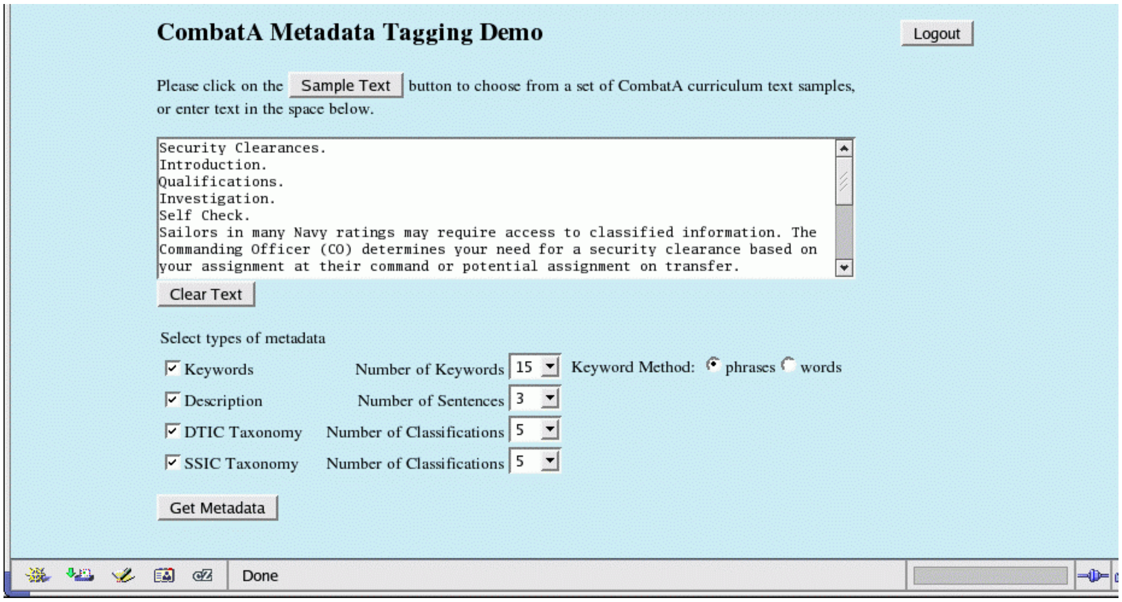
Sailors in many Navy ratings may require access to classified information. The Commanding Officer (CO) determines your need for a security clearance based on your assignment at their command or potential assignment on transfer. To apply for a security clearance, you must be a US citizen. Each Sailor needing a clearance will require an investigation. This investigation determines the Sailor's potential to protect information during the course of their duties. Once issued, a security clearance remains valid provided the Sailor continues compliance with personnel security standards and has no subsequent break in service exceeding 24 months.

The metadata tagging tool was a demonstration tool developed by Pearson Knowledge Analysis, Boulder, CO. The tool first allows the insertion of either a sample text or any text copied or typed into the tool. It also allows the selection of types of metadata such as keywords, sentence descriptions, and taxonomic classifications. Figure 2 below shows the initial interface for applying text and selected attributes for metadata tags.

Subsequent user interfaces display the content of the metadata such as the keywords and sentence descriptions generated as well as the taxonomic classifications selected. Prior to generating the actual XML metadata code, the interface allows the user to delete, modify, or add to any of the generated words, sentences or classifications.

Procedure

After reading a privacy act statement and completing an informed consent form and a brief demographic survey, participants were asked to read and provide 10 keywords that they felt best described the content of three separate paragraphs (i.e., 10 keywords per paragraph). Participants were told that the keywords could be either a single word or a word phrase.



CombatA Metadata Tagging Demo

Please click on the Sample Text button to choose from a set of CombatA curriculum text samples, or enter text in the space below.

```
Security Clearances.
Introduction.
Qualifications.
Investigation.
Self Check.
Sailors in many Navy ratings may require access to classified information. The
Commanding Officer (CO) determines your need for a security clearance based on
your assignment at their command or potential assignment on transfer.
```

Select types of metadata

☑ Keywords Number of Keywords 15 Keyword Method: ● phrases ○ words

☑ Description Number of Sentences 3

☑ DTIC Taxonomy Number of Classifications 5

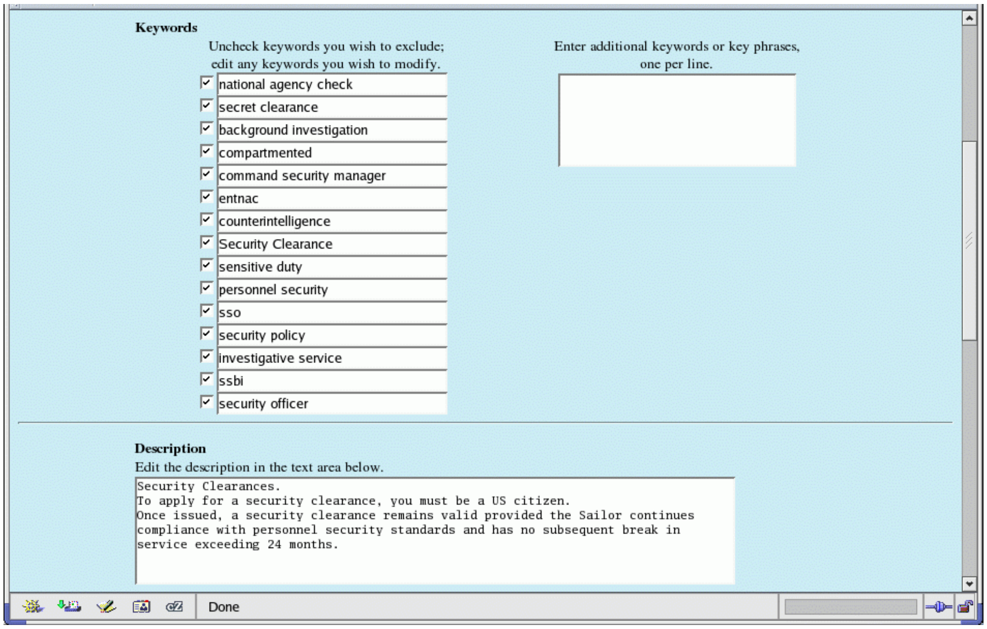
☑ SSIC Taxonomy Number of Classifications 5

Figure 2: User Interface of the LSA Metadata Tagging Tool

Once all three paragraphs were completed, participants were brought to a computer terminal and asked to use the automated metadata tagging tool to generate keywords, a three sentence description, and taxonomic classifications. The content used for this portion of the study was the first sample text for which participants had already provided key words.

Once the keywords were generated, participants were told to review the keywords generated by the tool and compare them to the original 10 keywords they had generated. They were then directed to modify the list as they believed appropriate by deleting any tool-generated terms they disagreed with, adding other terms as necessary. They were directed to keep the keyword list to 10 terms.

For each paragraph, participants were asked to rate the extent to which they agreed that the sentence summary and the taxonomic classifications accurately reflected the sample text.



Keywords

Uncheck keywords you wish to exclude; edit any keywords you wish to modify.

Enter additional keywords or key phrases, one per line.

☑ national agency check
☑ secret clearance
☑ background investigation
☑ compartmented
☑ command security manager
☑ entnac
☑ counterintelligence
☑ Security Clearance
☑ sensitive duty
☑ personnel security
☑ sso
☑ security policy
☑ investigative service
☑ ssbi
☑ security officer

Description

Edit the description in the text area below.

```
Security Clearances.
To apply for a security clearance, you must be a US citizen.
Once issued, a security clearance remains valid provided the Sailor continues
compliance with personnel security standards and has no subsequent break in
service exceeding 24 months.
```

Figure 3: Keywords and Description Interface.

Finally, participants were asked to perform the entire process of tagging a new text sample without any assistance from the facilitator. This included selecting the sample of text, setting the options for the number of keywords and taxonomic classifications, reviewing the generated information, and producing the XML formatted metadata. Following this tagging process, participants were then asked to provide ratings to statements concerning the usability of the tool.

No earlier than two weeks following initial data collection, each participant was asked to read and provide keywords for the third text sample they rated at the beginning of data collection. Follow-up data was collected in order to assess the degree to which individuals would agree with their own initial keyword generation.

RESULTS

Key Word Agreement

For each iteration of keyword generation, participants' lists were compared to all other participants' lists as well as the list generated by the tool for that sample text. An overlap score (agreement) was given for each comparison based on the number of terms that matched. The possible scores ranged from zero (0) to ten (10), with 0 indicating no terms matched to 10 indicating all terms matched. With only two exceptions, a match was scored only for identical terms. The exceptions included misspelled words or a difference as to whether the term was singular or plural.

The overlap scores for each sample text were averaged and are presented in Table 1. On average, participants matched just over one keyword for each sample text when compared with other participants.

Overlap was also scored within each participant assessing the number of terms from the initial keyword generation to the follow-up generation (Agreement with Self). On average, participants matched 4 items from their original list to their follow-up list.

Finally, participants' keywords generated from the first sample text were compared to the keywords generated by participants modifying the output of the computer tool (Text Sample 1 using Tool). In this case, the agreement score between participants averaged 5.5. Further, when participants used the tool and modified the keywords, the average overlap with the original keywords produced by the tool was 6.0.

Table 1: Agreement Means and Standard Deviations.

	Agreement with Other Participants	Agreement with Tool	Agreement with Self
Text Sample 1	1.38 (0.62)	1.10 (0.99)	
Text Sample 1 Modified using Tool	5.5 (0.52)	6.0 (1.49)	
Text Sample 2	1.49 (0.95)	1.10 (0.88)	
Text Sample 3	1.39 (0.77)	1.70 (1.25)	4.0 (0.94)

Usability

Participants rated a set of seven statements based on their opinion of the output of the software tool. Each statement was rated on a scale of one (1) to five (5), 1 representing "strongly agree" and 5 representing "strongly disagree." Table 2 presents each statement and the average score and standard deviation.

Table 2: Means and Standard Deviations for Subjective Questionnaire

Statement	**Mean (Std Dev)**
The three sentence summary accurately describes the paragraph.	2.7 (1.16)
The three-sentence summary fully describes the paragraph.	4.0 (0.47)
Of the 6 DTIC taxonomy classifications generated, the check box best reflects how I would classify this paragraph.	2.3 (1.49)
Of the 7 SSIC taxonomy classifications generated, the checked box best reflects how I would classify this paragraph.	3.1 (1.29)
The tool generated accurate taxonomy classifications.	2.3 (0.95)
The tool generates a sufficient number of potential taxonomy classifications.	2.1 (0.88)
The keywords I generated described the paragraph better than the keywords the tool generated.	3.0 (0.94)
There was ample range in the number of keywords that could be generated.	2.0 (0.94)

Participants also rated a set of five statements reflecting the usability of the tool based on their experience using the tool. Again, these statements were rated on a scale of 1 to 5, 1 representing "strongly agree" and 5 representing "strongly disagree." Table 3 presents each statement and the average score and standard deviation.

Table 3: Means and Standard Deviations for Usability Questions

Statement	Mean (Std Dev)
The software tool was easy to use.	1.5 (0.53)
The functions of the tool were understandable.	2.1 (0.99)
It was difficult to modify the keyword selections.	4.0 (1.41)
I understood how to use the tool.	1.8 (1.03)

DISCUSSION

The poor overlap of scores between the participants in this study is an example of what is referred to as the "vocabulary problem" (Furnas, Landauer, Gomez, & Dumais, 1987; Landauer et al., 2006). Thus, the problem associated with human tagging of an enormous amount of training material is not just the time and labor intensive nature of the task, but the subjective and highly variable output, as well.

For all three text samples used in this study, participant agreement scores averaged under two terms out of a possible 10 (i.e., less than one-fifth). However, when using the tool (and modifying the keywords based on their original judgments), participants showed a huge increase in agreement among each other: agreement scores increased from an average of 1.38 terms to an average of 5.5 terms. This even exceeds the agreement that participants had with themselves (i.e., an average of 4.0 terms) when asked to redo the task two weeks after the initial data collection. When modifying the terms using the computer, participants tended to rely heavily on the tool rather than keeping their original terms, demonstrated by the fact that following the use of the tool, participant's new lists agreed with the tool's original list on average of 6 terms.

It is of interest to note that the quality of the terms generated by participants was, in some instances, suspect. For example, four participants provided the term "requirements" to describe the first sample text. As a term used to locate a specific body of text, the term "requirements" is vague, at best, and could potentially represent a vast range of potential content. Other terms generated by participants that were similarly broad included "potential," "application," and "information."

Conversely, the tool generated very specific terms that often were not present in the actual sample text, yet represented strong relationship with the passage. For example, for the first sample text, a description of security clearance requirements, the tool generated the acronym DON CAF (Department of the Navy Clearance Adjudication Facility), the Navy's organization that grants security clearances.

Participants' responses to subjective questions of ease of use of the tool were quite favorable indicating the functions were understandable and easy to use. Participants' responses to questions of the utility of the tool were not as clear. While participants tended to agree that the tool provided an ample amount of keywords or taxonomic classifications, responses to whether they were accurate descriptors or whether they fully described the text were not as positive.

CONCLUSION

As shown in this study and demonstrated in previous research, there is very high variability in human agreement of naming (tagging) information. Arguably, training could lessen the variation between raters. However, even studies using domain experts (Furnas, et al., 1987), show poor agreement between raters. Allocating the function of tagging to a computer yields similar results. In sum, humans do not agree with each other nor do they agree with the computer.

There is a middle ground. This study found a marked increase in agreement scores when participants were able to use the terms generated by the computer, then modify them as they believed necessary. Whereas participants initially agreed with each other less than 20% when generating terms by themselves, they agreed with each other 55% of the time when using the tool to aid in keyword generation.

Ultimately, the true benefit of automated or semi-automated metadata tagging will only be known when used in the context of both tagging and, subsequently, searching. A very practical follow-on study might compare the hit rate for locating related material in a large repository comparing the use of automated, semi-automated, and human generated keywords.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the contributions of Dr. Thomas Landauer, Dr. Peter Foltz, Dr. Marcia Derr, and Mr. Wayne Hodgins for the development of the demonstration tool and their contributions to this research effort.

REFERENCES

Cooper, W. S. (1969). Is interindexer consistency a hobgoblin? *American Documentation, 20*, 268-278.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. _*Journal of the American Society for Information Science, 41*(6), 391-407.

Furnas, G.W., Landauer, T.K., Gomez, L.M., Dumais, S.T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM, 30,* 964—971

Gomez, L. M., Lochbaum, C. C., & Landauer, T. K. (1990). All the right words: Finding what you want as a function of richness of indexing vocabulary. *Journal of the American Society for Information Science*, *41*(8), 547-559.

Landauer, T, K (2002). On the computational basis of learning and cognition: Arguments from LSA. In B. H. Ross (Ed) *The Psychology of Learning and Motivation*, Academic Press.

Landauer, T. K., Foltz, P. W., Derr, M., & Leacock, C. (2006). Automated Metadata Content Tagging Software. (Final Report submitted under contract N61339-05-C09964). NAVAIR Orlando Training Systems Division, Orlando, FL.

Naval Personnel Development Command (2004a). Enabling the Navy revolution in training: An overview of the Navy Integrated Learning Environment (ILE). Version 1.1. October 2004

Naval Personnel Development Command (2004b). Initial capabilities document (ICD) for integrated learning environment information services architecture (ILE-ISA). Version 3.02 August 2004.

Tarr, D. & Barko, H. (1974). Factors influencing inter-indexing consistency. In *Proceedings of the American Society for Information Science 37th Annnual Meeting*, *11* 50-55.

Tinker, J. F. (1966). Imprecision in meaning measured by inconsistency of indexing. *American Documentation*, *17*, 96-102.