TECHNICAL REPORT

No. CCLS-07-02

Title: Selecting and Categorizing Textual Descriptions of Images in the Context of an Image Indexer's Toolkit

Authors: Rebecca J. Passonneau et al

Selecting and Categorizing Textual Descriptions of Images in the Context of an Image Indexer's Toolkit

Center for Computational Learning Systems

Technical Report: CCLS-07-02

Rebecca J. Passonneau
Center for Computational Learning Systems
Columbia University
New York, NY 10023
becky@cs.columbia.edu

Tae Yano
Department of Computer Science
Columbia University
New York, NY 10023
ty2142@columbia.edu

Judith Klavans
College of Information Studies
University of Maryland
College Park, MD 20742
jklavans@umd.edu

Rachael Bradley
College of Information Studies
University of Maryland
College Park, MD 20742
rlb@umd.edu

Carolyn Sheffield
College of Information Studies
University of Maryland
College Park, MD 20742
csheffie@umd.edu

Eileen Abels
College of Information Science and Technology
Drexel University
Philadelphia, PA 19104
eileen.abels@ischool.drexel.edu

Laura Jenemann
College of Information Science and Technology
Drexel University
Philadelphia, PA 19104
lj27@drexel.edu

ABSTRACT

We describe a series of studies aimed at identifying specifications for a text extraction module of an image indexer's toolkit. The materials used in the studies consist of images paired with paragraph sequences that describe the images. We administered a pilot survey to visual resource center professionals at three universities to determine what types of paragraphs would be preferred for metadata selection. Respondents generally showed a strong preference for one of two paragraphs they were presented with, indicating that not all paragraphs that describe images are seen as good sources of metadata. We developed a set of semantic category labels to assign to spans of text in order to distinguish between different types of information about the images, thus to classify metadata contexts. Human agreement on metadata is notoriously variable. In order to maximize agreement, we conducted four human labeling experiments using the seven semantic category labels we developed. A subset of our labelers had much higher inter-annotator reliability, and highest reliability occurs when labelers can pick two labels per text unit.

1. INTRODUCTION

We address the problem of information access, specifically of mining text for metadata for image access. Our study focuses on the development and evaluation of innovative methods to annotate text in order to extract high quality metadata.

Digital image collections range from informal, community based sharing of photos (Flickr) to formal, curated images of *primary sources and printed rarities* (New York Public Library, 480K images). Methods and means for indexing image collections with descriptive terms also vary widely. It has often been observed that search terms from image captions are likely to be insufficient for users' needs; the visibility of Google Image in leveraging the approach of searching by caption keywords and limited context is in part due to the vacuum it fills. Much remains to be learned about how descriptive metadata impacts image search, browsing of image collections, and other possible applications (e.g., sorting, classifying). There is suggestive prior evidence in a study that investigated the correlation of search terms with words from the topic titles and descriptions: users often created image queries using novel terms [4]. Manually adding descriptive metadata to existing image records that contain identifying metadata would have high costs; for books, Columbia University Libraries produced an estimate of $15.00 per volume for minimal descriptive metadata.

The work presented here was developed in the context of the Computational Linguistics for Metadata Building (CLiMB) research project, which has been investigating methods for automated support to image catalogers and other image professionals for locating subject matter metadata in electronic versions of scholarly texts [11]. The CLiMB project is developing a Toolkit for image catalogers that would allow them to access electronic versions of the types of texts they consult manually, but in a software environment that facilitates their work, and that could thus lead to improved access through richer indexing. Toolkit design goals include proposing terms for the cataloger to review. Because the Toolkit is co-evolving with changes in practice concurrent with the growth of

digital image collections, we have followed an iterative design process [16]. That is, we elicit feedback from image professionals, image catalogers, and end users along the way. Here we continue this approach in addressing the question of how to present Toolkit users with the most useful paragraphs for each image, and how to classify paragraphs or sentences into semantic categories that would be relevant for indexing. To address this goal, we have conducted a pilot study that examines paragraph preferences with respect to metadata quality, and a series of studies on human agreement in semantic classification of paragraphs and sentences extracted from electronic texts about images.

Our pilot study on rating paragraphs for metadata quality was a survey of indexing professionals to determine whether there is consensus on the types of paragraphs that image indexers would like to see, given an image and the task of selecting subject descriptor metadata from the paragraph. We presented subjects with images associated with pairs of paragraphs (image/text associatons) where we believed all paragraphs were at least somewhat relevant to the image. We expected to find a ranking of types of paragraphs, depending on a rough classification of the paragraph content. Instead, we found that in most pairs of paragraphs we presented, most indexers expressed a preference, and when they did so, they preferred the same paragraph.

The second study consisted of a series of experiments directed at developing a set of semantic categories for classifying paragraphs or sentences into a manageable set of relevant categories. Our goal was for the categories to yield a classification of the text extracts that image indexers would find relevant, and for the categories to be sufficiently robust that human annotators could be quickly trained to apply the labels in a consistent fashion. Two of the categories we developed were **ImgContent**, defined as text that describes the objective content of the image, and **Implementation**, text that describes the manner in which the depicted work was created, such as the style represented, or the technique used. The human annotators, or labelers, who participated in this series of studies were a combination of CLiMB researchers, and volunteers we recruited.

Our aims for the semantic labeling studies were inspired by other projects in the domain of scientific [22] [23] or legal [7] articles, in which human-labeled texts have been used as input for automatic text classifiers, with the goal of automated means to identify distinct regions within a scientific or legal text pertaining to semantic categories such as *claims, evidence, background*, and so on. We found that semantic classification of art history texts presents a much greater challenge than classification of legal or scientific texts, presumably because the content is more subjective. The categories for scientific articles, for example, reflect a highly standardized format that includes claims, supporting evidence, results, and comparison with other work. In particular, we found that measures of inter-annotator consistency varied quite a bit across our experiments, and sometimes were as low as those documented in more complex semantic annotation tasks, as in [9]. However, we also found that a systematic analysis of the dimensions we varied across our experiments yields very specific conclusions to explain the observed variation.

While each study produced somewhat unanticipated results, the combination has led us to a new model of how to present relevant text extracts to image indexers, and how to classify the text into categories that would provide more control over selection of metadata, and more potential for making use of controlled vocabularies and authoritative name databases. The external knowledge sources we have reviewed include the three Getty resources (Art and Architecture Thesaurus, Thesaurus of Geographic Names, Union List of Artist Names), the Library of Congress Authorities and Library

Historical Context Of the great projects built by Akhenaten hardly anything remains . . . . Through his choice of masters, he fostered a new style.

Implementation Known as the Amarna style, it can be seen at its best in

Image Content a sunk relief portrait of Akhenaten and his family (fig. 2-27). The intimate domestic scene

Historical Context suggests that the relief was meant to serve as a shrine in a private household.

Imgage Content The life-giving rays of the sun help to unify the composition and . . .

Key	
Historical Context	Social context (e.g., of use), or historical context of creation (e.g., commission)
Implementation	Explanation of a style or technique
Image Content	Content of the image or allegorical terms

Figure 1: Simplified illustration of semantic classification of text extracts

of Congress Thesauri for Graphic Materials, and ICONCLASS, a library classification for art and iconography developed by a Dutch scholar and supported by the Rijksbureau voor Kunsthistorische Documentatie (RKD); the image indexers who participated in our paragraph preference survey use these resources.)

Specifically, we conclude from the paragraph preference study that our initial selection of relevant passages from the art history survey texts was too broad, and that we should plan to build an automated classifier to select the more desirable paragraphs. The most important conclusions we draw from the five experiments where we collected human labelings on semantic classification of text extracts were: some but not all annotators can learn our semantic categories from our online, non-interactive training materials; by identifying the annotators with lower consistency rates, and allowing up to two labels per item, we can achieve high interannotator consistency. We also speculate that eliminating undesirable paragraphs is likely to lead to much higher interannotator consistency, given that we find marked variation in agreement levels across text extracts.

In section 2, we illustrate the aims of our two studies using a partial semantic classification of a paragraph similar to those used in our studies. In section 3, we summarize related work. Section 4 briefly describes the two art history survey texts we draw from, and the method we used for extracting image/text pairs for both studies reported here. Section 5 describes the paragraph preference survey, and section 6 describes four text labeling experiments. We discuss the results of the two strands of investigation in section 7, and what we deduce about the best way to proceed towards constructing automatic text classifiers for text extraction and categorization. We conclude in section 8 with general observations about the prospect for adding subject descriptors to large image collections, for linking those descriptors to controlled vocabularies, and for studying the impact on image search and image browsing.

2. BRIEF EXAMPLE

In the art history survey texts we are using, a single paragraph

or sentence can contain descriptive information about multiple images. Within a paragraph about a given image, the descriptive information can be categorized into distinct types which have a loose correspondence with categories of information discussed in the image indexing literature [4, 13, 2].

Figure 1 illustrates text from the first part of a primary layer associated with an image of a relief portrait of Akhenaten and his family. The image here is taken from the ARTstor Art Images for College Teaching collection (AICT): `http://www.arthist.umn.edu/aict/html/ancient/EN/EN006.html`. A subsequent paragraph (not shown here) compares the relief to another work by the artist, meaning that there would be some overlap between the text extracts associated with each image. The text fragment shown here has been separated into sentences that have been labeled to reflect three types of content. The glosses shown in the Key are highly abbreviated descriptions of a web document that defines and exemplifies the seven categories; an eighth category *Other* is available when none of the above applies.

3. RELATED WORK

In the literature reviews of the state-of-the-art in image search, image indexing, and use of controlled vocabularies for image classification and search, we have found repeated statements of the complexity of the problem [13], the need for empirical studies of a wide range of issues in image indexing [10], and the rapidly changing context, including evolving standards for image metadata [3]. In an early user study of a prototype Toolkit that involved a dozen image professionals [16], we found unalloyed enthusiasm for the project and an often expressed wish that we hand over a working model that could be used immediately, combined with an absence of clear specifications of what types of electronic texts we would need to handle or how we might recognize potentially useful subject matter metadata. Given the lack of existing tools to build on, and of standardized work processes and indexing practices among image indexers, we found that our first task would be to try to understand what image indexers could profit from in the general case, and to anticipate a new model for indexed image collections that would combine some of the features of controlled vocabularies with features of free form indexing terms and phrases.

There are relatively few discussions of inter-annotator or inter-indexer consistency for image indexing and classification tasks. Though there has been work on automated classification of text extracts [23] [7], including the issue of inter-annotator agreement in creating the training and evaluation data [22], no comparable study has been done for art history texts.

Studies of controlled vocabularies for image collections are relevant to our task, especially ones addressing the issue of relations among index terms. Shatford-Layne [13] suggested grouping the image indexing terms into four attributes (similar to our notion of category): *Biographical, Subjects, Exemplified*, and *Relationship*. Though we conceived our categorization scheme independently from her study, the two schemes are quite compatible. Her Subject attribute encompasses four of ours: ImgContent, Implementation, Interpret and Comparison. In the same paper, she also discussed a set of questions that such attributes (or categories) should address, such as the level of detail and the utility for end users.

Soergel [21] conducted a critical study of the Getty Art and Architecture Thesaurus (AAT). The pronounced absence of cross- references, which he points to as *the most serious shortcoming of the AAT hierarchy*, is perhaps the most relevant issue to our task. In the AAT, there are seven independent facets, comprising thirty three hierarchies; terms sit within a single hierarchy, despite the fact that in principle, a term might be relevant to multiple hierarchies or facets. We have observed the same issue in one of our experiments on classifying texts by their semantic content that restricted annotators to a single label: many annotators expressed difficulty in choosing a single label because some textual units seemed to belong in multiple categories. Soergel illustrated this point with many examples from Getty's thesaurus. He suggested that a *polyhierachical* system is more desirable for greater flexibility for end user's searching and browsing activities.

There are relatively few discussions of inter-annotator or inter-indexer consistency for image indexing and classification tasks. Two works that address the topic deeply and broadly are [14] and [8].

In the twenty some odd years since Markey's [14] comprehensive summary and comparison of forty years of inter-indexer consistency tests, no comparable review has appeared, and her observations still hold. Although her goal was to use the conclusions from previous work to sort through the issues involved in indexing visual material, all the tests referenced in her paper were on indexing of printed material. She notes significant variability among consistency scores. The agreement scores, using accuracy or percent agreement, range from 82% to a low of 4%. Markey sites a 1969 study by Zude that lists 25 factors that could affect indexing performance. Further, she presents her observations on factors she believes could increase inter-indexer agreement. She noted that greater inter-indexer consistency was attained when indexers employed a standardized classification scheme, comparable to a controlled vocabulary. However, even among the evaluations of indexer consistency with controlled vocabularies, the range is as wide as 34% to 80%.

The Giral and Taylor study [8] concerned indexing overlap and consistency on catalog records for the same items in architectural collections; they examined record data for title, geographic place names, and so on, including an analysis of subject descriptors. On large ($\geq$ 1400) samples of records from the Avery Index to Architectural Periodicals and the Architectural Periodicals Index, they compare proportions of items in their samples that match according to a variety of criteria, and compute 90% confidence intervals based on a formula for binomial proportions. Only 7% of items match entirely, and they find some element of overlap in descriptors in only about 40% of the remaining cases ($\pm$ 3%).

Percent agreement has the weakness that it is highly sensitive to the number and absolute frequency of categories assigned. If two categories are used, one of which is extremely frequent, percent agreement will necessarily be high [1]. While we use more robust methods for quantifying inter-annotator agreement, we find a similar range of values across four labeling experiments we conducted. We also point to several causes for this variation, and propose a means to control for it.

Markey [14] also examined the effects of indexers' familiarity with the classification scheme. Positive correlations between the two had been reported elsewhere, but her comparison of scores did not show a significant difference between the indexers with or without experience in using such schemes. Also, she found no higher levels of inter-indexer consistency among subject specialists, as compared with non-specialists.

On a somewhat exasperated note, Markey [14] stated that *the findings of inter-indexer consistency experiments appear to be inconsistent with one another, particularly in the evaluation of subject specialists, indexing aids, and indexer experience*. She goes on to observe that *non-subject specialists are capable of achieving a reasonable degree of consistency given clear, explicit directions and examples*. While this seems to suggest that most of the burden in achieving inter-indexer consistency is on training the indexer, our

own reading of her review is that there are other factors as well. For example, too little attention as been paid to the inherent difficulty of the materials being indexed. Markey herself noted two features of documents that affect inter-indexer consistency: the complexity of the document, which is difficult to quantify, and the document length, which is easy to quantify.

4. TEXTS

The domain of digital images and texts we focus on parallels the ARTstor *Art History Survey Collection (AHSC)*. ARTstor is a Mellon funded non-profit organization developing digital image collections and resources. The AHSC is a collection of 4,000 images that is the product of a collaboration with the Digital Library Federation's Academic Image Cooperative project and the College Art Association. One of our motivations for focusing on the AHSC is that it is based on thirteen standard art history survey texts, thus there is a strong correlation between the images and texts that describe them. The AHSC was designed *to include at least one image of every art object or monument reproduced in at least two of these standard survey texts*. The AHSC images all have metadata providing the name of the work, the artist, date, and so on, but very few have subject matter metadata. The few that do have subject matter metadata could be greatly enhanced by information extracted from the texts in the concordance.

We are currently using two of the texts from the AHSC concordance of thirteen art history survey volumes. Both books have a similar lineup of chapter topics, though there are some differences in text layout. They cover a broad time range, from Neolithic art to late 20th century; about one third of the chapters pertain to non-European regions. Each text contains roughly thirty chapters (approximately five megabytes in digital format). Each chapter consists of three to ten subsections, which in turn contain smaller subdivisions. The smallest subdivisions consist of about two to five paragraphs and each paragraph consists of about two to eight sentences. There are twenty to forty color images in each chapter.

For research purposes, CLiMB created electronic versions of the two texts, encoded in TEI compliant xml. TEI is a widely used interdisciplinary standard of text representation. The rules are defined in the TEI Lite customized schema. (See http://www.tei-c.org/Lite/teiu5_split_en.html for more detail of this schema.) Chapters, subdivisions, and paragraphs (but not sentences) have distinctive xml tags. The position where the plates appear is marked with an xml tag and the unique plate number.

To facilitate the construction of sets of image/text pairs for our survey and text labeling experiments, we employed software that had been created as a module for importing text into the CLiMB image indexer's Toolkit. The software module, which we refer to as RADIATE (for Random-Access Digital Image Archivist's Text Extraction tool), relies primarily on the relative position of xml tags for image plates, major text divisions, and paragraph boundaries. It takes a chapter as input, and produces a list of all the plates in the chapter, with each plate number associated with a sequential list of associated paragraph numbers. In an evaluation of its recall and precision against a standard we created manually, it achieved a recall of 0.69, precision of 0.75, for a combined F-measure of 0.72. Recall measures how many of the desired items were identified; precision measures how many of the items that were identified were correct; F-measure is the harmonic mean of recall and precision.

4.1 Creating the Sets of Image/Text Pairs for our Studies

We conducted labeling experiments with three sets of image/text pairs. Each set consisted of ten images and one or more paragraphs discussing the image directly or indirectly. All of the images and texts are excerpted from one of the AHSC art history survey texts. The images represent four distinctive chapters of art history; Ancient Egyptian art, Medieval art, Renaissance art, and Twenty Century Modern art. From those four areas, image/text pairs were chosen using a random selection process. We excluded any with more than six paragraphs. On average, a unit has 2.2 paragraphs (minimum 1, maximum 3), with 14.6 sentences (minimum 7, maximum 27), and 22.8 words in each sentence.

To generate the sets of image/text pairs, we used RADIATE for the first pass, then manually revised the output in order to exclude false hits, and to include paragraphs it overlooked.

5. PARAGRAPH PREFERENCE SURVEY

For the paragraph preference experiment, we were interested in the question of whether image indexers would have consistent preferences when asked to choose between two paragraphs about an image. If so, the survey responses could potentially provide criteria for the text extraction component of the CLiMB Toolkit. We were also interested in whether the preferences showed any pattern of correlation with the labeling categories applied to the paragraphs in question.

5.1 Survey Materials

We selected four works that each had at least four paragraphs associated with the work. Sometimes a work had multiple depictions (e.g., interior and exterior of an architectural work), so the full text extract of four paragraphs might discuss a different view of the same work. Each image came from a chapter representing a distinct time period: ancient Egyptian art, Romanesque art, Renaissance art, and postmodern European art. Using images available from free use sites, we created two questionnaire items per image. An image and the first two paragraphs of the text extract constituted one item, and the image with the second two paragraphs constituted a second item. This yielded a total of eight items.

The questionnaire included a variety of questions about the respondent's background, their reaction to the instrument, and so on. Here we discuss only the answers to the following two questions:

1. If you were told to select descriptive metadata to catalog the image from only one of the two following paragraphs, and your goal is to maximize the potential for user success and satisfaction in image searches, which paragraph would you pick?

2. On a scale of 1 to 5, with 1 being most desirable and 5 being least desirable, how would you rate each of the two paragraphs regarding how useful they are in providing descriptive metadata?

The subjects in the experiment were drawn from a set of half a dozen image catalogers who were participating in a separate cataloger workflow study conducted by the last three co-authors. The workflow study took place at the visual resource centers of three colleges and universities consisting of a mix of private and public institutions. The workflow study consisted of pre-interviews, observations, and post-interviews. Not all subjects who participated in the workflow study were able to complete the questionnaire. We have responses from four subjects per item, with the exception of two items for which we have three subjects (2.A, Image 2, first pair of paragraphs; 3.B, image 3, second pair of paragraphs).

5.2 Survey Results

Image Id	Item	Subject's Choices				Row Totals
		Both	Par 1	Par 2	Neither	
1	A	1	3			4
1	B		1	3		4
2	A			3		3
2	B		4			4
3	A			4		4
3	B	1	1		1	3
4	A		1	1	2	4
4	B		3		1	4
Column Totals		2	13	11	4	30

Table 1: Paragraph Preference Study, Question 1: Choose a Paragraph

Table 1 presents the survey results on question 1. Each row corresponds to a single item consisting of an image (1-4) and a paragraph pair (A or B); each column represents a selection users could make; the cell values give the counts for each selection on each item. There is a very strong pattern of preference for one paragraph over the other: in 24 out of 30 item responses, respondents had a preference. For six of the eight items, a majority of respondents preferred the same paragraph. In a subsequent section (subsection 6.4), we will discuss two metrics we use for quantifying the amount of agreement in datasets where we ask subjects to classify or label items we present, yielding data similar to that presented in Table 1. However, we cannot apply such metrics here because they require that the data represent the choices of the same set of subjects. While we have approximately the same number of subjects per item (three or four), they are not the same subjects in every row.

Regarding question two, very few paragraphs were rated *neutral* (6%), providing further support that overall, the image indexers in our study had strong judgements for and against paragraphs. Positively rated paragraphs (46%) were rated *somewhat desirable* (33%) a little more than twice as often as they were rated *highly desirable* (13%), and negatively rated paragraphs (47%) were rated *somewhat undesirable* (27%) a little more often than they were rated *highly undesirable* (20%).

6. TEXT LABELING EXPERIMENTS

6.1 Introduction

In tackling the problem of designing our semantic annotation scheme for classifying content of art history survey texts, we faced conflicting demands, and a wide range of possibilities. The criteria that in principle make for an ideal annotation scheme can be different from those that are best in practice. Even with respect to practical issues, there can be conflicting demands from the point of view of providing clear and consistent criteria for humans versus arriving at a classification for which automated classification techniques can be easily trained and evaluated. In this introductory subsection, we describe in general the conflicting demands we faced, the motivation for the five labeling experiments, and foreshadow the conclusions our results have led us to.

In the absence of an existing application with known specifications for classifying text, our goal in principle was to develop an annotation scheme with maximum coverage, meaning that the classes as a whole would cover all possible cases we observe in the two to six paragraphs associated with an image. This approach makes it possible to begin designing the text analysis methods concurrently with, or prior to, the cataloger user studies and end user studies that address the question of what types of information are more relevant and more useful for image searching and browsing. We can always discard categories we have identified that turn out to be of less relevance, but if we have not included the relevant categories in our analysis in the first place, we may lose an opportunity.

From the perspective of designing an annotation task, the ideal set of labels is one that can be learned quickly by the population from which the annotators will be drawn–meaning the annotators understand the criteria for each annotation label and feel confident about making the judgements involved–and for which an annotation interface can be easily designed so that data can be collected easily, quickly, and then stored in a convenient format for further processing. Thus it is better to require less training, to make use of fewer categories, and for annotators to have fewer steps to take in deciding on and applying a given category label.

The ideal classification scheme for machine learning is one that leads to classes that are easily distinguished by features that can be automatically extracted. In short, the more distinct the classes are, the better. At the same time, it is also preferable if the data to be classified is relatively homogenous, in the sense that selecting a different dataset to train on would lead to similar results.

Given the semantic complexity of the content expressed in art history texts, we could not satisfy the conflicting demands of maximizing coverage, minimizing the cognitive load for annotators, and creating optimal classes for machine learning. Instead, we conducted multiple annotation experiments as a means of exploring the tradeoffs. One of our goals was to design a set of categories that would *cover* most of the content expressed in the text extracts. We arrived at seven categories that we describe in the next subsection. We went through several iterations of guidelines that define and exemplify these categories in order to make them as understandable as possible for annotators, and also as relevant as possible to the image indexers and catalogers who consulted with us.

In our experiments to determine whether humans can consistently apply the categories, and under what conditions, we varied the number of labels that annotators could choose for a single unit, the relation of the labels to each other and to the labeled unit, and the size of the unit being labeled. Our experiments exhibit as wide a range of inter-indexer consistency as has been reported for related tasks, as noted in our literature review. By conducting a wide range of experiments on similar materials, we believe we have gained sufficient information regarding the nature of the task that we can account for some of the variation.

6.2 Semantic Category Labels

Over a period of four months, we developed a set of semantic categories for classifying paragraphs and sentences in the art history texts we are using, and detailed guidelines that define and exemplify the categories. Three criteria motivated the classification. Most important, the classes evolved by analyzing the texts; that is, we did not attempt to develop an independent set of categories based on existing image indexing work, and then see how they matched the texts. We took the information in the texts as our staring point. Second, the set of classes were designed to apply to most if not all chapters, and to allow most paragraphs or sentences to fall into a specific category, rather than to a default class, *None of the Above*. Last but not least, we worked with an image librarian at Columbia University and a metadata expert to arrive at a set they felt would be useful. (These two experts also served as beta testers of our training material.)

Table 2 summarizes the seven semantic category labels we currently use in our text labeling experiments. The column on the left indicates the name of the label, and the column on the right gives

Category Label	Description
ImgContent	Text that mentions the depicted object, discusses the subject matter, and describes what the artwork looks like, or contains.
Interpret	Text in which the author provides his or her interpretation of the meaning of the work.
Implementation	Text that explains artistic methods used to create the work, including the style, any technical problems, new techniques or approaches, etc.
Comparison	Text that discusses the art object in reference to one or more other works to compare or contrast the imagery, technique, subject matter, materials, etc.
Biographic	Text that provides information about the artist, the patron, or other people involved in creating the work, or that have a direct and meaningful link to the work after it was created.
HistContxt	Text describing the social or historical context in which the depicted work was created, including who commissioned it, or the impact of the image on the social or historical context of the time.
Significance	Text pointing to the specific art historical significance of the image. This usually applies to a single sentence, rather than to an entire paragraph.

Figure 2: Seven Semantic Categories for Labeling Texts

a highly abbreviated description of the type of textual content that should be assigned a given label. The labels appear here in the same order that they appear in the interface, which puts the most central category first (ImgContent), and which lists categories that have a similar focus together. Thus the first three categories are all about the depicted art work (form, meaning, manner); Biographic and HistContxt are both about the temporal and historical context.

During the first month, we arrived at a provisional set of six categories consisting of everything in Figure 2 apart from the italicized category, which now has the name *Implementation*, and developed our first set of guidelines. We added the seventh category after a month or so of pilot work. During the remaining three months we created three revisions of our labeling guidelines that define the categories. Our current guidelines give four pieces of information per category: the category label name, one or two questions the labeled text should answer, one or two paragraphs describing the category, and four examples illustrating an image and some associated text that the label applies to. For the ImgContent label, the questions are *Does the text describe what the art work looks like? What conventional use of symbols does the artist rely on?*

About halfway through the process, we began to create a sample of *training* cases that would allow labelers to practice applying the labels, and to compare their answers with consensus answers we collected during our pilot tests of labeling consistency.

The three months of revision included changes to the guidelines, and to the training examples. The types of changes were:

1. changing the category label name, e.g., to make the name easier to remember; to make the name semantically more general or more specific, so as to express a more central notion for the category;
2. revising the meaning of the category, e.g., by making it somewhat more inclusive, or somewhat less inclusive;
3. adding new examples, or revising existing examples;
4. expanding the definition to include the type of question the labeler should expect the labeled text to answer;
5. similar revisions to the training examples.

Exp	Set	Images	Units	Label Set	Labels/Par	Annotators
1	1	13	52	6	any	2
2	2	9	24	7	any	2
3	2	9	24	7	two	5
4a	3	10	24	7	one	7
4b	3	10	159	7	one	7

Table 2: Annotation Task Parameters

An example of the second type of revision is that our final definition of ImgContent includes iconographic content. While this is metaphorical (or allegorical in some cases), and was mentioned as a desirable category by our metadata and image cataloging consultants, we decided that the distinction between objective content and iconographic content was not always clearcut, would be less clear to labelers who did not have any background in art, and that it would be better to stick with fewer broader categories than to add further ones. We made this type of revision far less often than the other types. In about the third month, we added a question to every definition, and in the fourth month we continued revising the definitions. All revisions were based on input from three phases of beta testing with different test labelers.

The process of arriving at a set of semantic categories, clearcut definitions for them, and useful training examples that exemplified every category, was quite painstaking. We believe, however, that it was worthwhile. As we discuss in section 6.5, the first two co-authors did labeling for all experiments, but we expanded the set of labelers to five for experiment 3, and to seven for experiment 4. A comparison of all N combinations of labelers in both experiments (every set of two, every set of three, and so on) showed that among the seven labelers in our last experiment, two had very high consistency with other labelers across the board, and somewhat higher consistency than the first two co-authors. These two high-performing labelers had no prior experience. Thus, while three of the seven had very poor consistency with other labelers, the interannotator agreement results indicate that our guidelines and training materials are sufficient for at least some annotators.

6.3 Materials: Datasets, Annotation Constraints, Annotators, and other Task Parameters

We created three sets of image/text pairs for our experiments in human labeling of paragraphs and sentences, and we used them in the five experiments listed in Table 2. The second column of the table shows for each experiment which of the three image/text sets was used. Set 1 consisted of thirteen images with a total of 52 associated paragraphs. Set 2 consisted of nine images with a total of 24 associated paragraphs. Set 3, consisting of ten images taken from two new chapters, was used in two ways in the same data collection: 4a consisted of paragraph labeling (twenty four paragraphs), and 4b consisted of sentence labeling (159 sentences).

Throughout the experiments, the category labels continued to evolve, but the biggest change was between experiments 1 and 2, when a seventh category was added (see subsection 6.2).

Labelers were recruited from the CLiMB researchers and their acquaintances. Two who participated in all four experiments were the first two co-authors, who are a computational linguist and computer scientist specializing in Natural Language Processing, with experience in text categorization and similar areas. For experiment three, two of the three additional labelers consisted of two CLiMB project members with no prior text categorization experience, and the third was a doctoral student in art history. For experiment four, two of the five additional labelers were computer science students, but with no previous experience in NLP. The fifth was a student of landscape architecture with no previous experience in text labeling, a sixth was an image cataloging professional with an advanced degree in Library Science, and the fourth was an advanced undergraduate student who has participated in several NLP projects.

The two parameters of most interest for comparing the experiments appear in columns five (Labels/Par) and six (Annotators). For the first two experiments, the first two co-authors were the annotators, and the number of labels that could be assigned to a single paragraph was unrestricted. In experment 1, the maximum number of labels for a single paragraph was three; each annotator used three labels twice; 99% of the labelings consisted of one or two labels. In experiment 2, 71% of all labels from both annotators were one or two labels; the maximum for each annotator was four labels, which occurred once for each annotator.

Due to the relative infrequency of more than two labels in experiments 1 and 2, we added a restriction in experiment three that only two labels could be used. In experiment four, we restricted the paragraph level labeling further to a single label, but expanded the labeling task to include sentences. Our hypothesis was that for paragraphs where agreement would be contingent on multiple labels, we could compute the multi-label from the sentence labeling. Further, we hypothesized that we could identify a larger number of *pure* category instances, meaning a large majority of annotators agree on the same label, by the restriction to a single label.

For experiments 1 through 3, the labeling was done with pen and paper. For experiment 4, we implemented a custom browser-based labeling interface that included the guidelines, training materials, and labeling task. Labelers worked independently, at remote sites, and could suspend and resume work at will during the labeling task proper. However, labelers were required to go through approximately a one-hour training sequence prior to the first unit. For each image, paragraphs were presented one at a time. After selecting a label for the paragraph, the labeler would then be presented with the same paragraph in a sentence-by-sentence format, in order to label the sentences. After completing the paragraph and sentence labeling for a given paragraph, the next paragraph for the same image would be presented, until all image/text units had been fully labeled at the paragraph and sentence levels. At the end of each unit consisting of an image and all its paragraphs, labelers were given the opportunity to review and revise their choices, then again at the end of all units.

6.4 Evaluation Metrics

Recent annotation projects have tackled an expanding arena of linguistic and ontological features, and have developed increasingly complex annotation schemes to capture the distinctions of interest ([6, 19, 17]). As a consequence, there is growing need for research on the development of annotation schemes ([22] [18]), on novel approaches to measuring interannotator agreement ([1, 15, 20]), and on investigations of the relationship between measures of agreement and utility of the corpora (cf. [15]). In service of these long term goals, we have focused on developing an annotation language that can capture the distinctions of interest, possibly in variant forms, while maintaining high reliability along with high classificatory power, meaning that the classes cover the data well, and most annotators agree on when to apply them.

Our annotation scheme currently contains seven atomic elements, as described in subsection 6.2, above. As noted in the preceding section, we conducted four semi-controlled annotation experiments in which we varied experimental parameters, including how many labels an annotator could assign to the same unit (paragraph or sentence). For all experiments where we allowed annotators to assign multiple labels, the resulting label is essentially a set.

We report interannotator agreement using Krippendorff's α [12] and κ^3 [1], which range from 1 for perfect agreement to values close to -1 for maximally non-random disagreement, with 0 representing no difference from chance distribution. The two metrics differ in whether the estimation of expected agreement is derived from the observed distribution of values across all annotators, or whether a separate distribution is used for each annotator. The greater the difference in the α and κ^3 values, the more annotators differ in their distributions (annotator bias), as noted in [5]. Because annotators could make multiple selections, a distance metric for set-based annotations is used that gives partial credit when the annotators' sets overlap [15].

Benefits of agreement coefficients include factoring out agreement that would occur by chance, permitting comparison of results across different annotation tasks, and a rough measure of the difficulty of the annotation task [1, 5]. A disadvantage to agreement coefficients is that they do not provide a direct measure of what proportion of the data annotators agree on.

We also report mean F measure ($\overline{F}$), a metric based on recall (R) and precision (P) (see description in section 4). Because recall reports the proportion of desired items that is identified, and we have no gold standard for the items that should be identified (assigned a certain label), we take each annotation in turn as the standard, and compute recall of every other annotation against the proxy standard, and similarly for precision, then for F-measure, whose formula is:
$\frac{(1+\beta^2) \times P \times R}{(\beta^2 \times P)+R}$, using $\beta = 1$.

Thus $\overline{F}$ is the mean of the F measures that are based on taking each annotation as a proxy gold standard.

6.5 Results of the Text Labeling Experiments

6.5.1 Analysis of Agreement and $\overline{F}$ Values

Our global results for the four experiments appear in Table 3. Before we describe each experiment in detail, note that the two agreement metrics and the $\overline{F}$ value track each other fairly well, except in the case of experiment 4a (lines 4a and 4a'). This supports the utility of computing both metrics.

Experiment 2 has the best results: interannotator agreement is very high, indicating that the observed agreement is far greater than would be predicted by chance. The $\overline{F}$ measure is also quite high. If we could achieve these values when we do a large data collection to provide training data for machine learning, we would be in a very good position. The pros are that we would have high data consistency on two measures, we could use our existing tool for associating images and texts, and we would have maximum coverage of the input text. The cons are that the two annotators in this experiment (the first two co-authors, henceforth A1 and A2) would have to do all the labeling, which we might not have the resources for, and it would be somewhat hard to predict in advance how much data we would need to collect. In this dataset, the combinations of labels yielded twelve distinct values; a larger dataset might yield more. We cannot predict in advance how easy it would be to dis-

Exper.	Dataset	Variant	#Labelers	Alpha$_{MASI}$	Kappa$^3_{MASI}$	Mean F measure
1	Set 1	6/any	2	0.76	0.75	0.80
2	Set 2	7/any	2	0.93	0.91	0.87
3	Set 2	7/two	5	0.46	0.46	0.47
4a	Set 3	7/one	7	0.24	0.24	0.41
4a'	Set 3	merge 4b	7	0.36	0.36	0.62
4b	Set 3	one	7	0.30	0.30	0.43

Table 3: Interannotator consistency of paragraph labeling under multiple conditions

criminate among the twelve or more classes.

Experiment 1 differed from experiment 2 in two parameters: there was one less category, and a different (larger) set of units from two texts, rather than from a single text. Otherwise, the labeling criteria were the same: multiple labels could be assigned to each paragraph. We believe the improvement is due primarily to the increased experience of the annotators and a round of clarifications of the guidelines between experiments 1 and 2, rather than to the impact of the new label, or the different dataset.

Experiment 3 was the first time we used a larger set of annotators. We hypothesized that with each new annotator, the number of distinct multi-labels would increase, with the result that a large number of annotators would result in a large set of distinct classes, and corresondingly smaller classes. In order to guard against this possiblity, we restricted the number of labels that annotators could apply to a single paragraph to two. Recall (from subsection 6.3) that combinations of more than two labels were used relatively rarely. The resulting agreement and $\overline{F}$ values are less than 0.50, indicating poor results.

We attribute the decrease in data consistency at experiment 3 primarily to the relative unfamiliarity of a majority of the annotators with the labeling categories. The first two co-authors were among the five annotators, and the three others consisted of two members of the CLiMB research project from the University of Maryland, and a third who was a graduate student at Columbia University in the Art Department. The three new annotators were given a small *example set* that had already been labeled during experiment 1 or a pre-pilot. One of these (A3) received true *training* in that she had extensive feedback after labeling the example set and reviewing the consensus answers we had previously collected for the training examples. The other two (A4, A5) were shown the consensus answers but did not have extensive discussion comparing their reasons for selecting labels, and the contrast with the consensus labels.

Our justification for the speculation that the decreased labeling consistency is due to lack of familiarity with the labeling categories, and not the reduction in number of labels, is that when we computed interannotator consistency for all combinations of annotators from the set of five, we found that the two experienced annotators had values on the three measures (0.88, 0.88, 0.90) that were consistent with the results of experiment 2. A further observation regarding the role of experience is that every subset of three or four annotators in experiment 3 that included A3, the new annotator with the most training, had higher values than subsets without A3. For example, the subset (A1,A2,A3) had values around 0.70 on all measures compared with values around 0.50 for (A1,A2,A4) or (A2,A2,A5).

Experiment 4a yielded the poorest results. We attempted to improve the method of imparting the labeling guidelines to the annotators, and as described below, we have reason to believe we were successful. We attribute the poor results primarily to the highly restrictive constraint that annotators could only apply one label. Figure 1 illustrates why this might be the case: note that only part of a paragraph about the accompanying image is shown, and three labels have been applied (this is an actual example from our pilot data), including two instances each of HistContext and ImgContent. Also note that the sentence beginning *The intimate domestic scene* has been split into two pieces, each receiving a distinct label.

For experiments 4a and 4b, we used a labeling interface that was designed to insure that individuals would participate in some pre-task labeling of an expanded set of training examples. It was not genuine *training*, because labelers had only one opportunity to compare their answers with the consensus answers we had collected previously. The browser pages that presented their answers alongside the consensus answers we had previously collected provided textual explanations of some of the more subtle points. However, there was no opportunity for labelers to ask questions about the labels, and no iterative cycle that tested whether their answers became progressively closer to the consensus answers.

As in experiment 3, we computed interannotator agreement metrics for all combinations of annotators in experiment 4a. Our justification for the claim that the training materials (possibly including the interface design) show an improvement is based on the results for all 21 pairs of annotators. While values ranged from a low of 0.15 to a high of 0.32, there were three pairs of annotators with values between 0.29 and 0.32, with four of the annotators appearing in the seven pairs. The A1,A2 pair (the original annotators, with the most experience or training), had an interannotator agreement of 0.27. A1 appears in the second highest scoring pair, and A2 appears in the third highest scoring pair, but one of the new annotators appears in all three of the top three pairs. The subset of four consisting of the four individuals in the top three pairs had an interannotator reliability of 0.41 on both metrics.

For experiment 4a, we also computed interannotator agreement using a method of computing paragraph labels from the sentence level labels whose results are shown in the last line (4b) of Table 3. First, observe that the agreement and $\overline{F}$ results for 4b are quite similar to (somewhat better than) for the paragraph level results. We collected sentence level labeling in order to investigate the relationship between sentence level and paragraph level labels. We found, as expected, that the label assigned to a paragraph was very rarely distinct from any of the sentence level labels; this occurred less than 4% of the time. In 90% of cases, the paragraph level label corresponded to the sentence level label that was assigned to most of the sentences in the paragraph.

We also computed a paragraph level label for each paragraph from the sentence level labels. We created a relatively short label consisting of each distinct type of label applied to any sentence in the paragraph; if three sentences of a five-sentence paragraph were labeled ImgContent and two were labeled HistContext, the paragraph level label we compute is the multi-label consisting of ImgContent and HistContext. We also experimented with other combination methods, such as the union of all sentence level labels. However the best performing combination is the merge method using only the set of distinct labels, which is shown in the row 4a'.

Unit	α_{MASI}
1	0.40
2	0.24
3	0.25
4	0.03
5	0.02
6	0.17
8	0.38
9	0.12
10	0.12

Table 4: Agreement values (α_{MASI}) by image unit

The interannotator agreement improves substantially, though is still not at sufficiently high levels; the $\overline{F}$ value shows even more marked improvement.

Here we draw two conclusions from the results of experiment 4; in the next subsection on error analysis we draw a third. First, the comparison of the experiment 4 results of the best pairs of annotators with the best pairs of annotators in experiments 3 and 2 suggests that we could use multiple annotators who would have interannotator agreement as high or higher than the best performing pair by starting with a larger number of annotators, then filtering out the annotators with the poorest pairwise agreements. The possibility of identifying new annotators who perform well would mean that we have the potential to annotate larger amounts of training data without overburdening A1 and A2.

Our second conclusion is that we should revert to the labeling process of experiment 3 in which annotators were allowed to assign two labels. This would mean changing our labeling interface, the instruction set, and the training material. We anticipate that with these two changes, we could again achieve the high interannotator agreement and $\overline{F}$ values seen in experiment 2, and in the A1,A2 subset of experiment 3.

6.5.2 Brief Error Analysis

The dataset used in experiment 4 was set 3, which consisted of 10 images and 24 paragraphs. To understand in more detail the reasons for the low scores in experiment 4a, we looked at results for the nine individual units that had more than one paragraph; agreement metrics cannot be computed on single items. Table 4 shows the resulting α_{MASI} values, indicating a very wide spread. Units 4 and 5 had interannotator agreement that was at the chance level, while units 1 and 8 had agreement levels that were much higher than the overall agreement (0.38 and 0.40 as compared with 0.24). We can only speculate about the reasons for the low agreement on some units, which we will do in the following discussion section.

We also examined the distribution of numbers of annotators who agreed, to see if the labels differed regarding how often a majority of annotators selected the label. We found a striking difference across labels at the paragraph level only. At the paragraph level, all cases where four or more annotators agreed (a majority) on the same label involved one of three labels: ImgContent, Histcontext and Implement. At the sentence level, every one of the seven labels was used among the cases where 4 or more annotators agreed on a label.

7. DISCUSSION

7.1 Survey Results

The paragraph selection questionnaire indicates that image cataloger at three university visual resource centers, representative of the types of users we envision for the CLiMB Toolkit, generally have a clear preference for one of two paragraphs that discuss an image of an artwork. There are some paragraphs we presented them with that a majority find highly undesirable (paragraphs 2.B.2, 3.A.1 and 4.A.2). Of the remaining 13 paragraphs they saw, a majority agreed on the desirability of six.

A more detailed examination of the preferences revealed no clear pattern of ratings. Paragraph 1 of item 2.A was not selected, yet was classified by one respondent as *somewhat desirable*; this respondent rated both paragraphs equally, but selected one instead of both. Paragraph 2 of item 2.B was unanimously selected, but had as wide a spread of ratings as any item.

We began the paragraph selection survey after collecting labelings for only twenty image/text units, and from these twenty, there were only three that had four paragraphs associated with them; we created a new text/image unit for the survey, and had two of the labelers do this unit, which gives us a category label for all paragraphs used in the preference study. Due to the small set of units we could choose from, we could not systematically vary the category labels that had been assigned to the paragraphs in the preference study. Of the sixteen paragraphs, four were labeled ImgContent by all labelers, two were labeled Biographic by all labelers, and the remainder were assigned two to three labels covering all categories except Significance. We could find no pattern of association between the category label assigned to paragraphs, and the pattern of preferences or ratings.

7.2 Inter-annotator Agreement Results

The relatively wide range in agreement levels we see in Table 3 across all experiments is consistent with previous work, as noted in section 3. Earlier work that found relatively low interindexer consistency for indexing of visual materials [14] raises the possibility that different indexers might want different features in an indexing Toolkit. However, we believe our pilot study of paragraph preferences combined with our several experiments on labeling point the way towards a procedure for providing a component of an image indexing Toolkit that would present relevant paragraphs, and present a useful classification of those paragraphs.

First, given that image indexers at visual resource centers from three institutions all showed a clear preference for certain paragraphs associated with images, we conclude that we need to prune the set of paragraphs produced by our automatic selection process. We have not yet identified any criteria that characterizes the preferred paragraphs, but we could presumably build a classifier to sort paragraphs into the two binary categories, given that the human preferences are so strong. We believe it would require presenting the paragraphs we have already used (100 paragraphs in 3 datasets) to image indexers such that we collect four to five judgements per paragraph.

Second, from the results of our four experiments on labeling, we have identified two methods for maximizing interannotator agreement and $\overline{F}$ values: allow two labels per paragraph; start with many annotators, and prune out those who do not perform to threshold.

The error analysis in section 6.5.2 suggests two further possibilites. First, we could prune the set of labels to the three that most often receive majority votes. We would sacrifice coverage, but we might achieve more consistent labeling by doing so. Second, recall that that agreement is quite poor for certain units, and that the mean number of paragraphs per unit is just above two. If it were the case that many of the units contained one desirable and one undesirable paragraph, and that agreement is poor for undesirable paragraphs, then by pruning the paragraphs we might simultaneously

address the two goals of presenting Toolkit users with a smaller set of more consistently desirable paragraphs, and presenting Toolkit users with paragraphs that can be more reliably classified into relevant categories. While on the face of it, it might seem unlikely that the undesirability of a paragraph would correlate with difficult of achieving interannotator agreement, it is possible that some of the same factors are responsible for both. A paragraph that tries to cover more bases is less likely to be labeled with the same small set of labels (one or two) by most annotators; it might also seem too unfocused to be a rich source of metadata.

8. CONCLUSIONS AND FUTURE WORK

A key component to successful digital libraries is the creation of rich metadata for access, a process which is time-consuming and thus prohibitively expensive. We have demonstrated in our approach that using intelligent methods to extract small units of text from scholarly works will contribute to resolving some aspects of the metadata creation problem. Our results clearly indicate that we can achieve high interannotator agreement on markup of text for applying text mining techniques. We will test these hypotheses as we continue to embed our results in the CLiMB Toolkit and measure changes in cataloger and end user behavior.

The results of our two studies yield a specific approach for designing a text extraction and categorization module within the CLiMB image indexer's toolkit. We had designed our current text extraction module to include more rather than fewer paragraphs about a given image on the hypothesis that image indexer's would be interested in a more sources for metada. Due to the results of the paragraph preference survey, we plan to collect a large enough dataset of paragraph preferences that we can build an automated classifier to handle paragraph selection.

We also plan to collect a large dataset of paragraphs and sentences that have been categorized using our seven semantic labeling categories, in order to see if we can achieve success with a high coverage set of categories. We have identified three dimensions that need to be addressed in creating our training data. To handle the semantic fuzziness of many of the items to be labeled, the annotation needs to allow two labels per item, whether at the paragraph or sentence level. Given the infrequency of some of the labels that are somewhat less relevant, we may need to prune the set of categories. Finally, we discovered significant differences in our measures of consistency, depending on the paragraph being labeled, which suggests to us that we need to be more selective in our criteria for including textual units in our image/text association pairs. The labeled data will serve as training and evaluation data for a set of automatic classifiers.

In our future vision for the Toolkit, we hypothesize that when indexers select metadata from text that has been automatically classified into one of our semantic categories, it will be useful for the metadata to inherit the category assignment. Thus the output of the metadata selection process would be *typed* metadata. We believe this will facilitate automatic linkage of the metadata to controlled vocabularies that have similar classes, or whose classes have a definable relationship to ours. We hope this will lead to more flexible image search and browsing capabilities that can combine metadata consisting of short free form text selections with metadata using controlled vocabularies.

9. ADDITIONAL AUTHORS

10. REFERENCES

[1] R. Artstein and M. Poesio. Kappa^3 = alpha (or beta). Technical Report NLE Technote 2005-01, University of Essex, Essex, 2005.

[2] M. Baca, editor. *Introduction to art image access: issues, tools, standards, strategies*. Getty Research Institute, 2002.

[3] M. Baca. *Practical Issues in Applying Metadata Schemas and Controlled Vocabularies to Cultural Heritage Information*. The Haworth Press, Inc., 2003. Available through Library Literature, last accessed July 25, 2006. Also available through: http://is.gseis.ucla.edu/courses/Metadata/MBacaCCQ2003.pdf.

[4] H. Chen. An analysis of image queries in the field of art history. *Journal of the American Society for Information Science and Technology*, pages 260–273, 2001.

[5] B. D. Eugenio and M. Glass. The kappa statistic: A second look. *Computational Linguistics*, pages 95–101, 2004.

[6] D. Farwell, S. Helmreich, B. J. Dorr, N. Habash, F. Reeder, K. Miller, L. Levin, T. Mitamura, E. Hovy, O. Rambow, and A. Siddharthan. Interlingual annotation of multilingual text corpora. In *Proceedings of the North American Chapter of the Association for Computational Linguistics Workshop on Frontiers in Corpus Annotation*, pages 55–62, Boston, MA, 2004.

[7] B. Hachey and C. Grover. Sentence classification experiments for legal text summarisation. In *Proceedings of the 17th Annual Conference on Legal Knowledge and Information Systems (Jurix)*, 2004.

[8] A. Giral and A. Taylor. Indexing overlap and consistency between the Avery Index to Architectural Periodicals and the Architectural Periodicals Index. Library Resources and Technical Services 37(1):19-44, 1993.

[9] S. S. im Walde. Experiments on the automatic induction of german semantic verb classes. *Computational Linguistics*, pages 159–194, 2006.

[10] C. L. Jorgensen, 1999. Available at `http://digitalarchive.oclc.org/da/ViewObjectMain.jsp?fileid=0000002655:%000000059275&reqid=190`.

[11] J. Klavans. Using computational linguistic techniques and thesauri for enhancing metadata records in image search: The climb project. Article in preparation.

[12] K. Krippendorff. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills, CA, 1980.

[13] S. S. Layne. Some issues in the indexing of images. *Journal of the American Society for Information Science*, pages 583–8, 1994.

[14] K. Markey. Interindexer consistency tests: a literature review and report of a test of consistency in indexing visual materials. *Library and Information Science Research*, pages 155–177, 1984.

[15] R. Passonneau. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, 2006.

[16] R. J. Passonneau, D. Elson, R. Blitz, and J. Klavans. CLiMB Toolkit: A case study of iterative evaluation in a multidis ciplinary project. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, 2006.

[17] R. Prasad, E. Miltsakaki, A. Joshi, and B. Webber.

Annotation and data mining of the Penn Discourse TreeBank. In *Proceedings of the ACL Workshop on Discourse Annotation*, Barcelona, Spain, 2004.

[18] J. Pustejovsky, J. Castano, R. Ingria, R. Sauri, R. Gaizauskas, A. Setzer, G. Katz, and D. Radev. TimeML: Robust specification of event and temporal expressions in text. In *AAAI Spring Symposium on New Directions in Question-Answering (Working Papers)*, pages 28–34, Stanford, CA, 2003.

[19] J. Pustejovsky, P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, and M. Lazo. The TimeBANK corpus:. In *Proceedings of Corpus Linguistics*, pages 647–656, Lancaster, UK, 2003.

[20] A. Rosenberg and E. Binkowski. Augmenting the kappa statistic to determine inter-annotator reliability for multiply labeled data points. In *Proceedings of the Human Language Technology Conference and Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, 2004.

[21] D. Soergel. The arts and architecture thesaurus (aat): A critical appraisal. *Visual Resources*, pages 369–400, 1995. Available at `http://www.dsoergel.com/cv/B47_long.pdf`, last accessed July 31, 2006.

[22] S. Teufel, J. Carletta, and M. Moens. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the European Association for Computational Linguistics*, 1999.

[23] S. Teufel and M. Moens. Summarising scientific articles – experiments with relevance and rhetorical status. *Computational Linguistics*, pages 409–445, 2002.