CLiMB: Computational Linguistics for Metadata Building

Center for Research on Information Access
Columbia University Libraries

Contents

Appendices

CLiMB: Computational Linguistics for Metadata Building

Center for Research on Information Access
Columbia University Libraries

1. Executive Summary

The goal of the CLiMB project is to develop and assess the use of existing computational linguistic techniques as applied to the task of identifying and extracting rich descriptive metadata from text. The level of detail and the coverage which can be achieved by using automatic techniques would be far too costly to be achieved manually. The techniques to be developed in the CLiMB project thus offer the promise not only of improving the creation of descriptive metadata, but also of increasing access. Such metadata will be extracted from text which is in some way associated with an image, either explicitly or by topic. We will collect this metadata to explore its use for image collections. We propose a thorough and ongoing assessment of the metadata and an evaluation of its use within existing platforms. The Columbia University project team will be composed of an interdisciplinary group of librarians and computational linguists as well as research−oriented 'use experts' such as art curators, reference staff and selected faculty.

2. Goals & Overview: What will we learn? What new questions will be addressed?

The goal of CLiMB is to develop and test computer−assisted approaches to the creation of descriptive metadata for digital library special collections. The strategy proposed has the potential to provide rich, subject−oriented indexing for large collections that would otherwise be prohibitively expensive to describe and index using manual techniques. A further advantage of the approach set out here is that the descriptive metadata generated may be derived from authoritative scholarship in a way not normally feasible in standard cataloguing practice. CLiMB also promises to provide a platform for the development and testing of other innovative approaches to text−derived metadata generation and use that could lead in time to even more powerful search, retrieval and presentation tools for research and scholarship, including automatic metadata generation from non−English texts, linkages with subject thesauri, and cross−domain terminology mapping.

The premise of CLiMB is that content−based description of many digital collections often already exists implicitly in scholarly monographs and other published materials. The process of making this information explicit by distilling and linking it to metadata records could dramatically reduce the cost of providing effective retrieval for many digital library projects. In some cases such automatically−generated metadata might be the only affordable means of providing any type of content−based description.

CLiMB's strategy entails the use of standard tools and techniques from the field of computational linguistics as a starting point. Many of these tools are powerful but still require further development to be used for the high precision indexing task. Such tools will be evaluated, adapted and customized for use within the framework of current bibliographic description and existing electronic indexing and retrieval systems.

Our chief focus from the computational linguistic point of view will be to extend these currently available text analysis tools for the mining of descriptive metadata from prose descriptions of works of art. This descriptive metadata could be used either to enhance existing metadata or to

substitute for descriptions that might be prepared manually by specialist cataloguers. The strategy of building on existing computational tools and retrieval systems allows us to concentrate our attention on a) the specific types of linguistic analysis, filtering and post−processing that are needed to provide optimal content−oriented metadata, and b) developing targeted assessment methodologies to test the value of this approach within existing access platforms.

The CLiMB retrieval model assumes an environment in which a) digital representations of the collection (e.g., digital images, sound files, architectural projects) are or will be available online; and b) that the result of end−user searching and retrieval will normally be those representations themselves. It also assumes that published descriptions of the collection are available in scholarly monographs, textbooks, articles, exhibition catalogs etc. and that they that can be made available for "metadata mining" in order to build descriptive metadata records.

One of CLiMB's deliverables will be a set of recommendations for other projects for choosing candidate image collections along with texts. Our goal is both to develop methods and build tools to support those methods so that the wider community can generate CLiMB descriptive metadata for their own collections.

The CLiMB approach does not require that the target scholarly texts used for metadata extraction themselves be made available and viewable online, although where feasible this would provide an even richer knowledge base for many projects. This approach requires only that relevant scholarly documents be scanned and converted into temporary machine−readable texts for purposes of semantic analysis and metadata extraction. Once derivative data is built, the original source need not be retained nor displayed

3. Value and Justification: What Problems does the CLiMB Project Address?

As we have learned from collection−based digital library projects over the last few years, the most expensive and troublesome component is often metadata creation. Collections selected for reproduction on the web often lack cataloging at the item level, since the cost of such cataloging has usually been prohibitive. Even when funds are available, high quality cataloging of images, historical artifacts and other specialized collections requires the attentions of difficult−to−recruit specialist catalogers, and may take many months or years for completion. For these reasons digital collection cataloging is often limited to a very brief record, and retrieval options are thus limited to known item searching and browsing.

The nature of image cataloging necessarily varies depending upon the type and content of the images themselves and the institutional and historical context in which they are processed. Still, most approaches to image metadata creation recognize at least the following metadata subtypes:

1. basic identification (e.g., creator, date created, title, accession number or other unique identifier; sometimes 'cultural context')
2. technical description (e.g., capture details, reproduction techniques, physical media −of both the original image (if any) and the digital reproduction)
3. bibliographic context (e.g., related images, collection or publication information)
4. depiction (i.e., what the image is "of")
5. conceptual context (i.e., what the image is "about")

For example, the difference between describing what an image is "of" and what it may be "about" is illustrated in an example from the LC *Thesaurus for Graphic Materials:* A political cartoon depicting a basketball game in which the players are dribbling a globe may be "of" basketball but "about" international relations.

Effective digital library cataloging will normally give highest priority to documenting the formal aspects of description (1−3), since these address basic inventory, display and retrieval functions. Only afterward is content−based description (4−5) addressed, and it is in consequence often sparingly done with a caption or brief summary of what the image is "of."

A basic set of existing computational tools is already able to identify key words, phrases, names and dates related to individual images described in related texts. This capability is illustrated in the examples given below in Sections 6 and 7 and in the Appendices. On top of these robust existing techniques, we will build tools to assign relevance factors based on proximity analysis, frequency of occurrence, user−supplied stoplists or "booster lists," etc. Processing and comparing multiple descriptive texts for the same collection of images or objects may yield additional techniques for improving the relevancy weighting of terms so that users will be able to find a full range of images relevant to their needs.

Two further approaches to building high−quality metadata from programmatically generated descriptors will be explored in this project. We will test options for matching and correlating these descriptors against a) back−of−book indexes present in the textual works being processed; and b) published thesauri of image descriptors used in special collections cataloging, e.g.,

- ***LC Thesaurus for Graphic Materials***
- ***Getty Art & Architecture Thesaurus***

These approaches may allow us to assign more authoritative relevance weighting to specific terms and phrases; it may also allow the computer−assisted assignment of actual controlled vocabulary terms for use in standard, headings−based retrieval systems.

Successful strategies for computer−assisted metadata extraction would make it possible for many more scholarly collections to be made available generally to students and researchers. It also holds out the promise of bringing the best aspects of library and museum−oriented cataloging together with scholarly description and analysis, and making them both accessible within the same end−user access system.

4. Broader Impacts & Implications

As digital library initiatives evolve into deeper collaborations with scholars and researchers, the techniques proposed for CLiMB would have broad relevance to projects involving art objects, rare and specialized material, and museum and historical objects. Similarly, the development of techniques for the semiautomatic and automatic creation of descriptive metadata will affect the way planning and budgeting is performed for digital library projects. We believe that our results impact the entire process of creating and costing such collections. Another area of impact is on the way that image collections will become more navigable. Through the browsing and querying of rich and detailed metadata, which is not currently possible, we will begin to witness different types of search and retrieval behavior. This type of change was seen when the web was first used as the infrastructure for search and display. The use of web search engines is now standard methodology for most researchers and scholars, whereas just five years ago most people could not have envisioned the power of this resource.

Finally, the CLiMB project includes an extensive component for assessment and evaluation. Since we will be creating protocols for evaluation of access mechanisms that have never previously existed, we will contribute to the open literature on the way that people search, access, browse and navigate large amounts of descriptive metadata. We will publish and present our results at major conferences, and make our tools available as a way to disseminate these methods. This will enable the entire scholarly community to participate in evaluation at a larger level.

5. The Columbia University Context

Columbia University is in a unique position to build such an interdisciplinary project, requiring the expertise of librarians, subject specialists, and computational linguists. The Center for Research on Information Access (CRIA), directed by Judith L. Klavans, links natural language research in computational linguistics with the Libraries' goal of providing wide access to source materials. The purpose of the CLiMB project is to move towards this goal by means of text analysis for indexing. The Libraries are known for their extensive Digital Library program, with a focus on the Master Metadata File (MMF) for structural metadata. The goal of CRIA, located in the Information Services division of the University, is to create and build interdisciplinary research projects. The aforementioned natural language processing group in the Department of Computer Science at Columbia, chaired by Professor Kathleen McKeown, is one of the largest and most established in the country. Professor McKeown will be included in the CLiMB advisory team.

6. The CLiMB Project: Initial Results and Comparisons

As part of project planning for CLiMB, staff from Columbia Libraries and CRIA developed a proof−of−concept project in order to test certain overall project assumptions. These were:

a. that readily available computational linguistic software tools were effective enough "out of the box" in parsing narrative, descriptive texts that we could reasonably expect to be able to use and extend them to perform large−scale automated parsing and extraction of keywords and phrases from scholarly monographic and journal literature;
b. that we could envision strategies for adapting these software tools and refining the output such that highly relevant and meaningful vocabulary could be identified, filtered and weighted for use in metadata retrieval systems; and
c. that we could envision computer−assisted strategies for correlating extracted vocabulary with the specific individual works of art mentioned in source texts of varying type and style.

In this section we present our final results and conclusions from running the tests. Details of the input data, the processing steps, and output results are provided in Appendix A, B and C. The conclusions from these test runs gave us the confidence to proceed optimistically with the CLiMB research project.

For our initial testing, we chose two different types of images, with associated texts. (See Appendix F for descriptions of the image collections and descriptive texts we have targeted for the grant project itself.) Appendix A outlines the assumptions we made and the steps we followed. Appendix B presents the image we selected for analyzing text associated with Michelangelo's Bacchus (Michelangelo, Buonarroti, 1475−1564. Bacchus. 1496−1497. Marble. Museo Nazionale del Bargello, Florence, Italy) along with the text selection from the authoritative scholarly reference *Michelangelo* by Howard Hibbard (Harper and Row, 1974). Appendix C presents material from the architectural image collection of the architects Greene and Greene, along with a short text selection from the scholarly monograph *Greene and Greene* by Edward R. Bosley (Phaidon Press, Inc. 2000).

The computational tools available for text analysis have been developed largely for use in the analysis of news. In order to test the effectiveness of these tools in analyzing the specific type of scholarly, descriptive prose texts that would form the basis of this project, we selected from each marked−up sample a shorter section describing a single "work" (i.e., a Michelangelo sculpture and a Greene and Greene architectural project) for testing. These shorter sections were first parsed manually by a library staff member who was asked to list all noun phrases, keywords, personal names, geographic names, dates and references to other works, without any attempt to select those that might be more or less relevant to the work apart from filtering out

articles, etc. A combination of computational linguistic software tools, further described in section 7, was then assigned the same task. Adjusting for the slightly differing definitions as to what constituted a "noun phrase," the output of the computer−assisted processing was roughly equivalent to the manual output and with little effort could probably be made identical or even superior to the manual process. Both processes yielded a set of words and phrases that on their own constitute the beginning of a 'knowledge base' about the respective works. But while basic keyword or phrase searching of these raw descriptors might often yield interesting and useful results in a large image database, there would be clearly a great deal of retrieval 'noise' because of the inclusion of less relevant words and phrases and the unpredictable levels of direct connection to the work being described.

This helped us focus our research agenda on:

1. methods to identify those terms and phrases in which we could have higher confidence about their connectedness to the work being described (e.g., because of proximity, frequency or pattern);
2. methods to identify those terms and phrases that might be considered high−content−bearing in the specific subject or cultural domain of the work being described; and
3. ways in which search systems might handle retrieval and filtering of searches against descriptors at differing levels of relevance and connectedness."

The results of the exploratory study were positive and showed the strength of basic, readily−available computational linguistic text analysis tools. **Table 1** shows a comparison of results for the Greene & Greene text excerpt:

	Manual Method	CLiMB Tools	Correct Identification
Noun phrases	52	51	98%
Keywords	42	38	90%
Place Names	2	2	100%
Dates	5	5	100%
Non−place Names	15	12	80%
Related Targets	1	1	100%
Sum	117	109	93%

Table 1: Comparison of Manual and Computational Linguistic Methods

As can be seen in Table 1, the initial results of building metadata with these techniques are very strong. We identified from 80% to 100% of all targeted items, with an overall average of 93%. These results were obtained with no customization or filtering. Our software also found many additional words and phrases which will need to be evaluated for possible refinement of the process, or suppressed if they amount to textual "noise." At the same time, one of our research goals in the CLiMB project will be to work with the image librarian who will be responsible for identifying a set of gold standard terms and phrases, so that in the future we can compute standard precision and recall.

The initial study demonstrated the importance of our focusing the larger pilot on second−stage semantic filtering and metadata correlation; it also showed the clear need to develop a strong assessment component that could be used iteratively to help guide tool development and metadata strategies during the course of the larger project. Finally, the study revealed a number of significant additional research and procedural questions that will need early and focused analysis during the main project. For example, the results in Table 1 counted partial

overlap of noun phrase fragments as correct, but the question of how much overlap constitutes correctness remains to be determined. The proof will be in the user assessment for these phrases embedded within an access platform. However, our bottom line conclusion from running these tests was that the three assumptions we set out to test were positively confirmed.

7. Proposed Technologies

We propose three aspects to the technologies to be developed and delivered in this project. First, we will customize and extend existing computational linguistic tools. Second, we will embed the output of our results into existing platforms for assessment. Finally, we will contribute to the metadata development and standards effort.

7.1. Computational Linguistic Tools.

All linguistics software chosen for the full project will be freeware or freely licensable for research purposes so that future distribution of tools is not restricted. For the initial validation project we used three tools, each of which will be evaluated and perhaps extended as part of the CLiMB project:

a. Part−of−Speech Tagger: For the proof of concept study, we used a publicly available tagging tool provided by the Mitre corporation. (http://www.mitre.org/technology/alembic−workbench/) For the larger project, we will test different publicly available taggers to compare results and then select the best one for our purposes. The following sentence, taken from the Michelangelo text, illustrates the way the tagger marks up words:

"The flayed skin (probably not a tiger, but perhaps the legendary leopardus), full of grapes, with its head between the hooves of the little satyr, must symbolize life in death."

The result of tagging for part of speech adds the following information, shown in Table 2:

Word:	The	flayed	skin	…	with	its	head	between	the	hooves	of	the	little	satyr	etc
Tag:	Det	Adj	Noun		Prep	Poss	Noun	Prep	Det	Noun	Prep	Det	Adj	Noun	…

Table 2: Sample of Part of Speech Tagged Sentence

These tools are not 100% accurate, and one mistake can confuse the processing that follows. Therefore, accurate assessment of the impact of different taggers on accuracy will constitute part of our proposed work.

b. Noun Phrase Chunker: For the exploratory study we used a tool called LinkIT developed at Columbia (Wacholder 1998; Wacholder, Klavans, and Evans 2000; Klavans, Wacholder and Evans 2000; Wacholder, Klavans and Evans 2001). The output of the chunking creates keywords and noun phrases such as those shown in Table 3. In Table 3 we also show the difference between phrases that the manual method identified compared to CLiMB techniques:

	Manual	**CLiMB**
flayed skin	flayed	skin
tiger	no	yes
legendary leopardus	no	no
grapes	yes	yes
hooves of the little satyr	no	Little satyr
life in death	life, death	life, death

Table 3: Noun Phrases and Keywords

For the full project we will test several noun phrase 'chunkers' to see which performs the best for this task (Ramshaw and Marcus, 1995; Argoman, Dagan and Krymolowski, 1999; Brill and Ngai, 1999; Ngai and Yarowsky, 2000). We will also use different ways to view noun phrases, for example, simplex noun phrases as in “life”, “death” or complex phrases such as “life in death.”

7.2. Image Retrieval and Display Systems

Our plan is to test the search and retrieval functionality of CLiMB−generated metadata in several different search systems. The primary object of this is to identify and implement a single demonstration system that will allow effective and iterative assessment and comparison of the CLiMB approach to metadata generation. A secondary objective, however, will be to examine the ways in which CLiMB−type metadata behaves in some key digital library search environments.

The initial set of test systems will be a) the Luna Insight product, b) the local library catalog at Columbia or another partner site, and c) the Columbia "Master Metadata File" (SQL / IBM's DB2).

The following are among the questions we will explore:

- Is it better to load large numbers of generated descriptors with varying degrees of relevance & confidence levels, or a smaller number that have only the highest levels of relevance/confidence?
- Is it desirable to encode and search machine−identified proper names and geographic places as such, or is it more effective to search all CLiMB output as keywords?
- Do existing search systems display CLiMB−type metadata results in such a way that the results can easily be understood, manipulated and navigated?
- What are the kinds of factual and research questions that could be answered by standard Keyword/Boolean searches against a single collection using CLiMB generated metadata? Against multiple collections?
- How much does search and retrieval improve when generated metadata is matched, filtered and weighted with a back−of−book index or with generated metadata from another text on the same collection?

If time allows, we would also like to do some preliminary investigation within the Columbia Master Metadata File environment into ways of optimizing retrieval and "data mining" on a knowledge base generated by CLiMB processes. Unfortunately, current indexing and retrieval tools available for digital library projects are generally at an immature stage and rarely go beyond low−level keyword/Boolean techniques.

It is generally recognized that the indexing and display requirements for library−assigned metadata differ substantially from those needed in a full−text display and indexing system. Since CLiMB−type metadata may in fact occupy a midpoint between traditional metadata and full text, approaches that combine elements of both search models may improve retrieval. To explore this we will identify indexing and retrieval tools at Columbia and elsewhere that may be adaptable for use in this project and/or build on our own local indexing tools. In particular we will be attempting to identify functionality such as:

- keyword−in−context result displays
- more powerful and research−oriented result set manipulation

- more interactive & interoperable browsing & searching modes
- interactive and automatic query optimization
- concept−based retrieval
- graphical content mapping
- the use of visualization techniques over interconnected phrases

The retrieval and display challenges raised by a large CLiMB−type metadata knowledge base parallel the overall challenge of navigating large−scale digital libraries and could have implications beyond this project.

7.3 Image metadata standards

CLiMB's content−oriented descriptors will be easily accommodated in any of the standard image description data formats, either as keywords or uncontrolled vocabulary. If standard thesaurus terms are effectively derived from the CLiMB descriptors, they would be candidates for "controlled vocabulary" fields.

The following are relevant data element standards with an indication of where CLiMB metadata might be encoded, if needed:

- **VRA CORE 3.0:** VRA Core Categories, Version 3.0. a project of the Visual Resources Association Data Standards Committee (http://php.indiana.edu/~fryp/vracore3.htm)

  SUBJECT Description: Terms or phrases that describe, identify, or interpret the Work or Image and what it depicts or expresses. These may include proper names (e.g., people or events), geographic designations (places), generic terms describing the material world, or topics (e.g., iconography, concepts, themes, or issues). Data Values: recommend AAT, TGM, ICONCLASS, Sears Subject Headings

- **Dublin Core:** Dublin Core Metadata Element Set, Version 1.1: Reference Description (http://dublincore.org/documents/dces/)

  Subject and Keywords: The topic of the content of the resource. Typically, a Subject will be expressed as keywords, key phrases or classification codes that describe a topic of the resource. Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme.

- **USMARC:** MARC 21 Concise Bibliographic (http://lcweb.loc.gov/marc/bibliographic/ecbdhome.html)

  650 − SUBJECT ADDED ENTRY−−TOPICAL TERM (R)

  653 − INDEX TERM−−UNCONTROLLED (R)

7.4 Text Conversion Requirements

Although the focus of CLiMB is on developing, testing, and distributing computational linguistic methods, in order to achieve our goals we will need to convert some existing texts into digital form. Note that this step is not an explicit goal of CLiMB, but rather one step in the CLiMB process. As noted, all converted texts for which we do not obtain permission to display will be destroyed once our techniques have been applied. All metadata will have the source text attributed.

The testbed monographs and descriptive texts will be processed through an optical character recognition (OCR) system and thus converted into plain text (ASCII) versions of the originals. Such automatic conversion often produces inaccuracies, thus producing "dirty" OCR output. For

the CLiMB project, believe that "dirty OCR" introduces too many uncertainties into the process to be able to test adequately the specific techniques and hypotheses of CLiMB. This is especially applicable when comparing CLiMB metadata to existing controlled vocabulary metadata and to back−of−the−book indexes where high precision is required. To accurately and reliably prove that CLiMB is successful on its own terms, relatively clean text input is important.

Our current estimated pricing for converting texts into minimally accurate ASCII texts, with minimal structural markup is between $75−$100 per 100 pages of text. This estimate is based on our recent experience with text conversion projects.

8. Collaboration with Related Projects

As mentioned in Section 7.3, the CLiMB project for building descriptive metadata has natural links with several ongoing projects. We have already seen such potential in the early development stages of this project. For example, we have discussed potential for linking the kind of metadata that we will build for images with the Cornell Fedora project (http://www.cs.cornell.edu/lagoze/). We would like to explore additional links with other projects such as ArtStor, the Indiana image repository, and the DLF academic image exchange project. The selection of image collections and of source descriptive texts will be performed in the first part of the first year of the project so that we establish our partners in early stages. We will also draw on collections at Columbia which will permit us to test our technologies. (See http://www.columbia.edu/cu/lweb/projects/digital/about.diap.html), including the Digital Aviador project.

We will also explore working with the SDARTS group at Columbia as part of CliMB. SDARTS (http://sdarts.cs.columbia.edu/default.html) is a protocol for metasearching over online document collections, and was developed as part of the Digital Libraries Project at the Columbia University Department of Computer Science. The purpose of SDARTS is to make a wide variety of collections with heterogeneous interfaces accessible under one uniform interface

9. Rights & Permissions

The purpose of this project is to develop techniques to create fuller access to images, but not necessarily to present source texts themselves. In fact, as noted in Section 2, the CLiMB approach does not require that the texts used for metadata extraction themselves be made available and viewable online. Our project requires only that relevant scholarly documents be scanned and converted into temporary machine−readable texts. These on−the−fly conversions are not saved but rather are used in a one−time process for the building of descriptive metadata. Once derivative metadata is built, the results will be incorporated into standard access platforms as part of indexing similar to extensive manual keyword creation.

10. Methodologies for Testing and Evaluation

This section outlines the two types of evaluation that we will perform as part of the CLiMB project. The first concerns the accuracy of the descriptive metadata derived with computational linguistic tools, while the second describes the use of the metadata for search and access within existing platforms for scholarly purposes.

A preliminary description of collections and texts to be used as testbeds for this project may be found in in Appendix F. Because of the breadth and depth of Columbia's collections in art and architecture the majority of these texts are available to us on site. As discussed below, we will make adjustments to this selection as needed as we learn more during the course of the project.

10.1. Evaluation of Software and Computational Tools

In the past decade, enormous progress has been made in the computational linguistic analysis of natural language. The development of fast and robust tools for identifying parts of speech, phrases, and phrasal expansions has enabled the development of a host of related applications never before possible, such as summarization, translation, and mapping from natural languages (like English or French) into SQL−compliant databases. However, the process is far from perfect, and each application brings along with it certain requirements that involve tailoring generic tools. The metadata extraction application that we propose in the CLiMB project also brings its own set of unique requirements.

We propose two levels of evaluation for tools in the CLiMB project. The first involves the comparison of available tools in order to decide which ones best suit the CLiMB goals. The second involves judging the accuracy of results as we iteratively tailor the tools to fit the task. Each of the tools we decide to use must be evaluated in light of the CLiMB application. Furthermore, by using a set of tools which are pipelined together, we must ensure that each step is as accurate as possible in order to avoid propagation of errors down through the system.

For example, in Section 6 we explained the steps we followed for the CLiMB exploratory project, followed by Section 7 in which we discussed the tools we used. Each of these tools will be evaluated along with at least one, and sometimes several, alternatives. The criteria for evaluation will be accuracy over speed, since accuracy and precision is of top importance for the CLiMB application. For other applications, e.g. topic spotting, the opposite holds. In this case, speed would be more valued than precision. In the ideal world, both speed and accuracy are required, but given the state of the art, there is usually a tradeoff and such choices must be made in light of the application.

At the same time, since output is often faulty, we will evaluate the nature of the these errors and develop customizations to compensate for any errors which are important to the application. Some noun phrases are identified incorrectly, and we will need to trace the source of these errors. This could be due to the fact that most taggers are trained on newspaper text so the usage of words might tip choices in a direction which is not applicable to our language type.

On a more subtle note, our exploratory project showed that breaking noun phrases into their smallest parts is not always desirable. In the example in section 6, we will need to decide how to break up phrases with the preposition "of" as in "the hooves of the little satyr" or "full of grapes." These decisions are typically made on noun classes, so identifying the classes of nouns that apply to the image metadata application will be part of the evaluation.

The data in Appendices B and C illustrate some of these errors clearly. We have not cleaned this output yet, and it is clear that some of the preprocessing and post−processing required will become obvious as we customize the tools. At the same time, other changes will only become known as the technology development team interacts with the scholarly user group for feedback.

10.2. User−Oriented Testing & Evaluation

The core of the CLiMB project is the use of computational linguistic techniques over trusted scholarly texts associated with image collections for the extraction of rich metadata. The image collections that we will use will include a range of image types (e.g. photographs, paintings, three−dimensional objects) and will also cover a range of collection type (e.g. grouped objects vs. individual objects). The testbed size will vary widely, and what is needed for the collection of useful metadata will be determined iteratively as we proceed. For example, the size of the texts associated with an individual image, with a set of images on a single object, and with a collection of objects will vary. Indeed, certain very popular objects have extensive text, which could be mined for metadata (e.g. the Mona Lisa), whereas other objects have carefully linked narrow associated texts (e.g. the Chinese gods collection). Examining the impact of this variability in

text size and type on the quality of our results will be reported as one of our findings. We will balance our choices and will remain open to re−balancing as our results emerge during the course of the project.

At the same time that we test with users, we will work to determine the minimum for achieving a critical mass of useful automatically−identified metadata. In fact, among our deliverables will be the information for other projects on what size text and what type of text is most useful for what kinds of collections. This is the kind of issue we will report on as part of the reporting structure for this project that we have outlined. We will produce a written account of our developments and results for Mellon at least three times per year during the duration of the CLiMB project.

Defining a "Successful Search." A key task of the project will be to develop a set of working standard against which to measure effectiveness and level of user−satisfaction. This would necessarily have to be customized for a project environment in which:

- retrieval is performed against computer−derived enriched vocabulary and phrases, rather than thesaurus based terms or raw keywords;
- existing indexing and retrieval software packages are employed, rather than a system optimized for this specialize type of metadata;
- the user will be searching in a relatively narrow subject−specific environment rather than in a heterogeneous OPAC or art or cultural materials database.

In carrying out this task, we will begin by reviewing search and retrieval functionality and effectiveness. We will test both in selected commercial and noncommercial art−related databases and in existing implementations of the three proposed test systems (see 7.2 above) with our target user groups to try to develop a set of model, "successful" searches.

Flexible Text Selection & Iterative Prototyping. After we have a working standard of 'success' we will need for each image collection to evaluate and select texts that appear to be best suited to metadata extraction −− assuming there are multiple parallel texts to work from. There is, naturally no particular need for randomization or balance in this selection; a project goal is to determine whether there are sufficiently "good" scholarly texts to produce "good" metadata.

After a preliminary working standard of success is in place, we will then begin iterative prototyping and testing to experiment with different balances of a) initial manual and computer−assisted identification of works described; b) initial text review to identify specific stop words and phrases, applicable external thesauri, user−supplied 'booster' terminology; c) the value added by back of book indexes; d) the value added by multiple texts corresponding to the same images (or projects or sites).

Testing Scenarios. One of the goals of the project is to proceed to carefully evaluate each of the variables, so that by the end of Year 2 we will have a set of solid and reliable results. This will require the kind of iterative prototyping described in the bulk of the proposal. Testing, development and prototyping would proceed according to scenarios such as the following:

1. **Select approximately ten existing collections of varying sizes**.

   Our criteria will be that half (about five) will be related in some way (e.g., time period, artist, medium), and the other half will be dissimilar collections. For example, at least one of the collections will be sculptural (e.g. of carvings or religious art.) Another criterion will be that each collection must already be online and have at least one complete and authoritative scholarly text on the collection itself. If possible, we will prefer more than one text, although for certain collections we realize this will not be feasible. A final desired criterion is that at least some collections have been formally cataloged including content description at the item level, so we can try "gold standard"

testing. This final requirement is more difficult to meet, but one which will be important to seek in order to fully evaluate our techniques against manual cataloging.

2. **Test different combinations of variables, for example:**

    a. Create metadata test datasets from:
        - 1 collection, 1 scholarly text, no back−of−book index
        - 1 collection, 1 scholarly text, with back−of−book index
        - 1 collection, 3 scholarly texts, no back−of−book index
        - 1 collection, 3 scholarly texts, with back−of−book index
        - 2 collections, 2 scholarly texts, no back−of−book index
        - 2 collections, 2 scholarly texts, with back−of−book indexes

    b. Run tests against each dataset in:
        - the Columbia online catalogue, CLIO
        - Luna Insight
        - the master metadata format (MMF) being developed at Columbia

    c. Develop user assessment scripts to test:
        - objective retrieval precision recall against "gold standard"
        - subjective curator/librarian retrieval success
        - subjective scholar/researcher retrieval success
        - subjective non−scholar retrieval success

    d. As part of c, include such search strategies as:
        - known item searching
        - fuzzy known item searching
        - search by known style / technique
        - search by fuzzy style / technique

This testing scenario would cycle through: subjects (depictions), subjects (topical and other), historical context, colors, dates, locations and names. We will design the scripts to take maximum advantage of the proprietary search retrieval and display functionality of each individual system.

Broader Questions and Considerations. One a more general level, CLiMB may yield important feedback on broader questions and issues such as:

- What types of research needs will CLiMB metadata support most effectively?
- How important is it to scholars and researchers to be able to have available the type of content−based retrieval CLiMB will allow? Are there different user groups who have significantly different search and display needs that would affect the design of, e.g., a customized CLiMB−based retrieval systems.?
- How can "relevancy to the user" be defined in an enhanced metadata search & retrieval system, as opposed to a mechanically assigned relevance of hits in a particular result set?
- Can the inevitable "noise problem" of false and less relevant hits resulting from the CLiMB process be reduced to sufficient levels that users will not become frustrated with the system? What are the specific types of false hits that are the most troublesome (e.g., the "not" problem as in "'Night' and 'Day' are _not_ carved to fit the sarcophagus lids as their opposites are...") and how difficult or expensive would it be to address them?
- Can the tradeoffs between cost and effectiveness and user satisfaction be understood clearly enough to allow for the development of a calculus for determining the best balance of preliminary text markup, manual post−processing of extracted metadata, and effective user discovery, retrieval and display within the context of a larger and growing metadata knowledge base?

- How well can existing retrieval and display systems for images and other digital library objects support effective searching against large image collections with enriched metadata records−−and possibly many other types of related information? Or will a new generation of retrieval systems and interfaces need to evolve for this?

11. Project Deliverables

Project deliverables will include:

a. The creation of a demonstration system with sufficient critical mass to allow effective search & retrieval of targeted digital collections using a combination of manual and machine−built metadata;
b. The development of an evaluation technique for the generated descriptive metadata;
c. The creation of a training and test set of manually created 'model descriptions' which will be made publicly available and against which incremental progress can be measured;
d. An assessment by scholars, librarians and other relevant user communities of the effectiveness of using this type of machine−generated metadata for retrieval and research;
e. A set of software tools, guidelines and procedures that would allow other institutions to freely use these same techniques with their digitized collections;
f. A set of criteria and recommendations for choosing candidate image collections and parallel descriptive texts for use by others implementing the CLiMB approach;
g. A set of recommendations for the minimal set of manually−assigned cataloging elements needed to effectively 'anchor' computer−generated metadata descriptions (e.g., title, unique identifier or accession number;
h. An assessment of possible additional research and development tasks needed for further enhancement of these tools and broader application of the technique.

12. Project Timetable

YEAR 1

1. **Project Preparation**

    a. Recruit and hire project art librarian, programmer and other staff (See Appendix E)
    b. Finalize research agenda with partners, advisory board and project art librarian;
    c. Evaluate and select specific existing computational tools;
    d. Define additional basic software toolsets that need to be developed as part of the project;
    e. Select and purchase/operationalize image database & retrieval system for project use;
    f. Refine selection of test image collections and associated texts for initial testbed;
    g. Convert relevant printed texts to ASCII and mark up to identify works/sites/projects

2. **Develop & Test Linguistic/Metadata Tools**

    a. Optimize and extend existing software tools for parsing and sorting keywords, phrases, etc.;
    b. Develop strategy and create tool for differentially identifying and optimizing for "associational context" of works described in narrative text;
    c. Develop additional software metadata extraction tools as needed;
    d. Develop customizable, parameter−driven tool for filtering of different result sets;
    e. Investigate options for computer−assisted identification of "works";

f. Compare computer−assisted metadata with cataloger−generated metadata in controlled sample.

3. **Create End−User Demonstration System**
   Configure, populate and optimize selected database retrieval system to support assessment phase.

4. **Conduct User Assessment 1**
   Develop assessment model & tools; conduct preliminary assessment testing.

5. **Prepare Interim Reports.** Prepare interim reports on progress and results (Dec., May, Aug.). Develop mechanisms for informing colleague institutions about the project and encouraging feedback.

YEAR 2

1. **Refine Tools & Processes**
   Using feedback from preliminary formative user assessment, adjust, modify and enhance metadata generation & filtering process; if feasible, replicate demonstration system in alternate database software with different indexing, retrieval and interface functionality. Test back−of−book index filtering & matching.
2. **Conduct User Assessment 2**
   Conduct additional assessment testing on revised demonstration system(s).
3. **Prepare Economic Analysis**
   Develop alternative cost models showing tradeoffs between manual editing, computer−assisted metadata extraction, manual post−processing, further tool development.
4. **Externalize Tools & Processes**
   Package software customizations and project tools for external use, using open−source model; document and publish "cookbook" of procedures and options for other sites to replicate and build on project achievements.
5. **Prepare Final Report & Recommendations**
   Prepare and submit final report that includes possible future research agenda for building on current project, including possible new models of scholarly research tools that combine generated metadata, digital images & objects & other scholarly texts and databases; user−driven metadata enhancement techniques, etc.

In Year 2, we will also prepare a follow−on project with feedback from our advisory board and partners, which we envision might include such future work as:

1. **Thesaurus Matching**
   Develop strategies for matching & correlating back, external thesauri, gazetteer/geographic lists
2. **Test Multiple Text Correlation**
   Develop semantic and statistical strategies for using multiple texts to enhance relevancy and scope of generated metadata
3. **Consider extensions to multilingual data**.
   Run exploratory tests of bilingual mapping software over text associated with images to validate approach for multilingual indexing.

13. Project Team

This section provides an overview of staffing to show the team members. We propose a tightly integrated team consisting of image use specialists, digital library specialists and computational linguists. Full details of responsibilities with description of staff members is given in Appendix E.

The project will be managed by Judith L. Klavans, Director of the Center for Research on Information Access in the Columbia Universities Libraries. A project team meeting will be held bimonthly with full staff to review progress and ensure integration. Since this is an interdisciplinary project, such meetings are key to smooth progress. If funded, we will build a small and focused external advisory board consisting of experts in metadata structure, scholars in related fields who will be our targeted user group, technical experts in image digital library projects, and applied computational linguists.

We are requesting funding for:

1. One project manager and computational linguistics researcher – 20%
2. One digital collections/art research librarian who will be fully dedicated to this project – 100%
3. One programmer at 50% for transferring research tools to existing platforms, building assessment tools, performing test runs, and analyzing results
4. One full−time 12 month Ph.D. level graduate student from the Department of Computer Science to develop and test computational linguistic technologies
5. Supervisory time for a Digital Projects Implementation Librarian at 10%
6. Funding to cover about 50% of a project associate who will be responsible for collection of progress reports from the various participants in order to prepare quarterly dissemination of results, and to integrate our progress with related projects on metadata and on image metadata.

In addition to funded positions, a Senior Art Librarian, Angela Giral, Director of the Avery Architectural Fine Arts Library at Columbia, will be involved at 5% of her time. She will take responsibility for overseeing hiring of the image librarian and ensuring that all image−related standards are incorporated into the project. She will also be involved in selection of image collections as we chose partners and collections in Year 1. Finally, she will participate in evaluation by contacting art curators, reference staff, and selected faculty. The internal advisory group will also include the University Librarian, Patricia Renfro, and the Director of Bibliographic Control and Processing, Robert Wolven.

We will request travel for the external advisory board to meet on an annual basis, which could also include an associated scholarly event.

Further detail and description can be found in **Appendix E.**

14. Conclusion

The CLiMB project at Columbia University proposes innovative approaches to the identification and extraction of descriptive metadata for images. Our techniques will apply known computational linguistic tools to text associated with these images. By developing applications to filter and refine current output, we will identify sets of related terms in context. These terms will be incorporated into existing platforms that are being used for image access, such as the Luna system. In this way, the use of this rich and robust data by scholars can be assessed in a controlled way. Our goals are to impact the way that image collections are indexed and thus browsed and accessed. Our preliminary tests have shown that such techniques provide promise for vastly improving access. The only step left is to develop and test the tools and techniques. This final step will be achieved within this CLiMB proposal.

15. Bibliography

Alembic Workbench Project. http://www.mitre.org/technology/alembic−workbench

Argamon, Shlomo, Ido Dagan and Yuval Krymolowski, *A Memory−Based Approach to*

Learning Shallow Natural Language Patterns. Journal of Experimental and Theoretical Artificial Intelligence (JETAI), volume 11 (3), 1999.

Brill, Eric and Grace Ngai, "Man vs. Machine: A Case Study in Base Noun Phrase Learning", in: *Proceedings of ACL'99*, University of Maryland, MD, USA, 1999.

Evans, David K., Judith L. Klavans and Nina Wacholder (2000) "Document Processing with LinkIT".*RIAO 2000, Recherche d'Informations Assistee par Ordinateur*. Paris, France, pp. 1336−1345.

Klavans, Judith L. (1989) ``Computational Linguistics,'' in *Contemporary Linguistics: An Introduction.* Mark Aronoff, ed., St. Martin's Press, New York.

Klavans, Judith L., Nina Wacholder and David K. Evans (2000) "Evaluation of Computational Linguistic Techniques for Identifying Significant Topics for Browsing Applications", in *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*. Athens, Greece.

Grace Ngai and David Yarowsky, "Rule Writing or Annotation: Cost−efficient Resource Usage for Base Noun Phrase Chunking", in *Proceedings of ACL−2000*, Hong Kong, 2000.

Lance A. Ramshaw and Mitchell P. Marcus, "Text Chunking Using Transformation−Based Learning", in *Proceedings of the Third Workshop on Very Large Corpora*, Cambridge, MA, USA, 1995.

Wacholder, Nina, Judith L. Klavans and David K. Evans (2000) "Evaluation of Automatically Identified Index Terms for Browsing Electronic Documents", in *Proceedings of the Joint Conference on Applied Natural Language Processing and the North American Chapter of the Association for Computational Linguistics (ANLP−NAACL)*. Seattle, Washington, pp. 302−308.

Wacholder, Nina, David K. Evans and Judith L. Klavans (2001) "Automatic Identification and Organization of Index Terms for Interactive Browsing", in *Proceedings of the First ACM/IEEE−CS Joint Conference on Digital Libraries (JCDL)*. Roanoke, Virginia, pp. 126−134.

Wacholder, Nina (1998). "Simplex NPS sorted by head: a method for identifying significant topics within a document," *Proceedings of the COLING−ACL Workshop on the Computational Treatment of Nominals, Montreal, Canada, August 16, 1998.*

---

16. Proposed Budget [provided separately]

Appendix A: Background Material for CliMB Exploratory Study

1. Overview

As part of project planning for CLiMB, staff from Columbia Libraries and CRIA developed a small pilot project in order to test certain overall project assumptions, namely:

a. that readily available computational linguistic software tools were effective enough "out of the box" in parsing narrative, descriptive texts that we could reasonably expect to be able to use and extend them to perform large−scale automated parsing and extraction of keywords and phrases from scholarly monographic and journal literature;
b. that we could envision strategies for adapting these software tools and refining the output such that highly relevant and meaningful vocabulary could be identified, filtered and weighted for use in metadata retrieval systems; and
c. that we could envision computer−assisted strategies for correlating extracted vocabulary with the specific individual works of art mentioned in source texts of varying type and style.

---

2. Methodology

Our methodology was to select two scholarly texts describing architectural projects and art images, respectively. Then:

- Portions of the texts were scanned and converted to a simple TEI format
- Text samples were then manually marked up to flag the location of all target works
- Within each of the two samples, a contiguous subsection relating chiefly to a single 'target object" was subjectively identified
- From these contiguous selections, keywords, phrases, names, places, dates and references to other works (members of the target class) were manually extracted to use as a point of reference. Extraction was systematic, though prone to human error; only basic filtering was done to remove articles, prepositions, etc.
- Repeated occurences of each word or phrase were tallied & recorded
- On the same selections, a standard semantic parser and part of speech recognition tool was run; keywords, phrases, places, non−place names, dates and references to other works were listed and tallied.
- The output from the manual and machine−assisted parsing was then programatically compared and statistics generated on retrieval recall and precision.

To process the sample texts, we used existing software tools that had been built under NSF funding in Columbia's Center for Research on Information Access and Computer Science Department.

The stages of automatic markup illustrating steps of this process included:

1. Input text was labeled for part of speech (e.g., noun, verb, preposition). For the pilot, this was done with a publicly available part−of−speech tagger called "Alembic." For the larger project, we will test diffet publicly available taggers to compare results and then select the best one for our purposes.

2. The tagged sentence was then passed through a noun phrase analyzer. For the pilot we used a tool called LinkIT developed at Columbia (by: Wacholder 1998; Wacholder, Klavans, and Evans 2000; Klavans, Wacholder and Evans 2000; Walcholder, Klavans and Evans 2001). For the full project we will test several noun phrase 'chunkers' to see which performs the best for this task. Any software chosen will be freeware or freely licensable for research purposes so that future distribution of tools is not restricted.

3. The final step performed by LinIT is to group noun phrases in various ways, for example:

The tools that we used create what are called simplex noun phrases, which is the smallest type of noun phrase. It is clear, however, that we will require more complex phrase identification for the larger project, optimized for subject domain.

It is also clear that we will need to sort and group noun phrases in ways that metadata retrieval systems and scholarly users will be require. It is precisely this type of tool extension and customization that we propose to accomplish during the full project.

3. Working Terminology

a. **"Target Text":** The text to be processed for the extraction of metadata
b. **"Target Class":** A defined set of entities, objects or concepts being described in the "Target Text;" e.g., Greene & Greene building projects mentioned in Ted Bosley's book; paintings mentioned in the book "Western Art Since 1500"; musical compositions mentioned in the "Norton Anthology of Western Music: Classic to Modern".
c. **"Member of Target Class/Target Object"**, e.g., a single building project such as "L.A. ROBINSON HOUSE (PASADENA, CALIF.)"; a single artwork such as "Guernica"; a single musical composition such as "Beehoven's Kantate auf den Tod Kaiser Josefs II".
d. **"Associational Context"**, e.g ., the extent of text surrounding an incidence of a Member of the Target Class deemed likely to yield relevant associated words and concepts; e.g., "from 100 words before to 200 words after an incidence"; or "from the start of the paragraph preceding a significant incidence until the end of the paragraph in which the next significant incidence of a *different* Member occurs."

---

4. Working Assumptions

a. The Target Class for performing 'CLiMB' analysis will be determined by the project client in advance; normally the client will also provide an authoritative list and/or unique names or numbers for all Members of the "Target Class" expected to appear in the text to be analyzed along with a forumulaic description of ways in which references to Members may appear (e.g., "Robinson House" *or* "Laurabelle Robinson's" etc.; "Picasso's Guernica" etc.; "Beethoven's Kantate auf den Tod Joseph II" *or* "WoO 87" etc.)
b. The textual location of each incidence of a Member of the Target Class will normally be flagged in advance, either manually or through a separate computer−assisted means that is not the primary focus of this phase of 'CLiMB'.
c. 'CLiMB' analysis ideally incorporates a flexible approach to assigning high or low "Significance" to individual incidences of Members of the Target Class appearing in the Target Text, e.g., based on density of incidence of a single Member (such as number of mentions in a single paragraph); or proximity of incidences of different Members (e.g., in a brief listing or summary of projects). The rules for determining significance should be applicable in such a way as to allow subsequent processes to be performed only on those incidences that meet the desired threshhold of significance.
d. 'CLiMB' analysis will entail an initial determination of the presumed "Associational Context" to be applied within the Target Text, i.e., the extent of text surrounding a significant incidence of a member of the target class within which a useful degree of relevance may be inferred. The Associational Context will need to be customizable according to the nature of the Target Class and the nature of the Target Text itself; for this reason a simple & flexible set of rule−based options for defining Associational Context should be available and adjustable based on iterative review of 'CLiMB' results.
e. Within the Associational Context of an incidence, 'CLiMB' analysis should be able to identify words and word groups, such as: noun phrases, individual noun key−words, personal names, place names, other proper nouns, dates; in some cases, external thesauri or word lists may need to be brought to bear to enable accurate identification, e.g., a separate list of major US place names; 'CLiMB' analysis should also be able to correlate and infer relationship between

different Members of the Target Class appearing in proximity to one another;

f. 'CLiMB' analysis should provide an easy mechanism for the client to specify a "stop list" of words, phrases, names, etc. that are _not_ to be considered relevant to the analysis and that should always be excluded from analysis and output. Ideally there would also be a mechanism for the client to specify a "hit list" of words or phrases that are automatically to be given extra weight & significance in assigning relevance.
g. One type of output of a specific 'CLiMB' analysis would be a listing of each unique Member of the Target Class accompanied in each case by a labeled listing of associated noun phrases, keywords, names, dates, places and related Members, arranged where feasible by presumed relevance based on attributes such as frequency, or proximity.

---

5. Agenda for Further Analysis & Testing

a. **Back of Book Indexes.** What added value or functionality could "back of book indexes," when available, provide in terms of filtering results or providing semi−controlled vocabulary?
b. **Geographic Authority Lists.** Would it be possible to improve the automatic identification of geographical names within a given text by matching on external gazetteers or authority lists?
c. **Multiple Texts per Target Class.** How much added value might accrue from using more than one text describing the same target class as a way of validating, filtering, clustering or weighting extracted metadata?
d. **Automatic Work Identification.** Can a machine−assisted technique be made effective enough to automatically identify described individual members of the target class? What type of easily−applied iterative, adjustable matching tool could be created for this purpose?
e. **Indexing & Retrieval.** Will standard relevancy−based indexing & search engines be able to adequately retrieve on 'CLiMB'−type metadata? What intermediate metadata markup or weighting might be needed to improve search engine precision?
f. **End−User Tools.** What interface−based tools would be needed for a user to review, sort, select and filter the results of a 'CLiMB'−based search, assuming the likelihood of large result sets for many words and phrases? Do these toolsets exist currently, or would they have to be developed?
g. **Associational Context.** Is it possible to generalize about effective "associational context" (see definition above), or perhaps categorize typical sets of parameters based on the type of text, to allow for computer−assisted determinations? Or must the 'CLiMB' toolset include a flexible content analysis tool to allow customization of the associational context for each text?
h. **Post−Processing & Editing.** How much benefit is there in post−processing extracted metadata, e.g., to remove common words or filter against a stop list? Can this be done using a set of standard parameters or is it necessary to customize this for specific domains or individual texts?
i. **Costs and Trade−offs.** Can a cost−effective mix of manual and automated processing be identified that would make it feasible to provide metadata for large collections of images at 'reasonable' cost? Would basic image cataloging provided by a cataloger be better, worse, the same or just different from the output provided by a 'CLiMB' approach?

Appendix B: Michelangelo Sculpture Image & Text Used for CLiMB Exploratory Study

Detail

Detail

Michelangelo, Buonarroti, 1475−1564. ***Bacchus.*** 1496−1497. Marble. Museo Nazionale del Bargello, Florence, Italy.

---

1. Text Processed for CLiMB Study

[From: Hibbard, Howard Michelangelo. 2nd ed. New York : Harper & Row, 1985.]

Messer Iacopo Galli, a Roman gentleman of good understanding, made Michelangelo carve a marble <anchor id="1">Bacchus, ten palms in height, in his house; this work in form and bearing in every part corresponds to the description of the ancient writers − his aspect, merry; the eyes, squinting and lascivious, like those of people excessively given to the love of wine. He holds a cup in his right hand, like one about to drink, and looks at it lovingly, taking pleasure in the liquor of which he was the inventor; for this reason he is crowned with a garland of vine leaves. On his left arm he has a tiger's skin, the animal dedicated to him, as one that delights in grapes; and the skin was represented rather than the animal, as Michelangelo desired to signify that he who allows his senses to be overcome by the appetite for that fruit,

<pb n="39">

15. Maerten van Heemskerck, view of the sculpture garden of Jacopo Galli, Rome. 1532−5

16. Michelangelo, <anchor id="1">Bacchus (detail)

<pb n="40">

17. Michelangelo, <anchor id="1">Bacchus (detail)

<pb n="41"> and the liquor pressed from it, ultimately loses his life. In his left hand he holds a bunch of grapes, which a merry and alert little satyr at his feet furtively enjoys.

Michelangelo's first masterpiece [14] was carved in 1496−7 from Riario's block and at his expense. Perhaps quite soon it found its way into the collection of Riario's friend and neighbor Jacopo Galli, where it can be seen looking like one of the antiquities, its right hand broken off, in a drawing of the early 1530s [15]. Perhaps because it was always planned as a free−standing statue, Michelangelo carved a figure that is unusual in his work; from a frontal position the pointed base and raised cup deflect the viewer to the right: the chief view is shown in illustration 14 − but the composition begs to be seen from several points of view around 180 degrees, from front to back. This slow movement is encouraged by the fascinating torsion of the coy little satyr, which also furnishes the support needed by a standing marble statue [17]. Michelangelo's figure is standing in one of the traditional art−poses of antiquity, but seems to sway back tipsily as he eyes his large cup, mouth open.

Vasari, writing about what we would call the transition from Quattrocento to High Renaissance art, emphasizes the beneficial influence of antiquity, citing the newly−discovered 'appeal and vigor of living flesh' and the free attitudes, 'exquisitely graceful and full of movement.' This new spontaneity, 'a grace that simply cannot be measured', and the 'roundness and fullness derived from good judgement and design' are perhaps seen here for the first time in modern sculpture. In addition the statue is novel in its depiction of the god of wine, naked and enraptured with his own sacred fluid. Michelangelo combined familiar ancient proportions with a suspiciously naturalistic rather than ideal nude body. Although several figures of Bacchus survive from antiquity, none is so evocative of the god's mysterious, even androgynous antique character: as Condivi says, it is in the spirit of the ancient writers. Nevertheless, grapes, vine leaves, a wine cup, a skin, and a little satyr can all be found accompanying one or another of the ancient representations.

The <anchor id="1">Bacchus is at first disconcerting. We imagine the sculptors of antiquity producing noble, heroic works; when we think of sculpture by Michelangelo, the David or Moses perhaps spring first to mind [25, 107]. Here we have instead a soft, slightly tipsy young god, mouth open and eyes rolling [16], his head wreathed in ivy and grapes, as pagan and natural as Michelangelo could make him. Since the statue was in the open for over half a century its polished surface is weathered.

Jacopo Galli, a banker, was the intimate of a Humanistic circle that included not only Cardinal Riario but also such men as the writer

<pb n="42"> Jacopo Sadoleto, whose dialogue Phaedrus was set in Galli's suburban villa. We can therefore suspect that Michelangelo was given learned iconographical information to incorporate into his statue. The teacher of Bacchus was Silenus, who was reputed to be the father of the Satyrs. The flayed skin (probably not a tiger, but perhaps the legendary leopardus), full of grapes, with its head between the hooves of the little satyr, must symbolize life in death. The ancient cults of Dionysus−Bacchus were associated with wine and revelry but also with darker things: grisly orgies, ritual sacrifice, the eating of raw flesh. Some of this veiled frenzy seems to have been incorporated in the attributes of the <anchor id="1">Bacchus, and a sense of mystery filtered down even to the naive Condivi. In later years Michelangelo returned to the image of a flayed skin as symbol of his own plight, both in poetry and in the eerie figure of St Bartholomew in <anchor id="30">The Last Judgement [163].

In a letter of 1 July 1497 Michelangelo wrote his father:

Do not be astonished that I have not come back, because I have not yet been able to work out my affairs with the Cardinal, and don't want to leave if I haven't been satisfied and reimbursed for my labor first; with these great personages one has to go slow, since they can't be pushed...

This means that the <anchor id="1">Bacchus was finished, but obviously it did not lead to further commissions from Riario, who was not attracted by modern antiquities. A further letter of 19 August reports that

I undertook to do a figure for Piero de' Medici and bought marble, and then never began it, because he hasn't done as he promised me. So I'm working on my own and doing a figure for my own pleasure. I bought a piece of marble for five ducats, but it wasn't a good piece and the money was thrown away; then I bought another piece for another five ducats, and this I'm working for my own pleasure. So you must realize that I, too, have expenses and troubles . . .

Michelangelo's complaints are made at least partly in response to his father's; the older man was threatened with a lawsuit following his brother's death. But perhaps we can also detect a genuine unhappiness, which Michelangelo could not analyze, and to which he referred in later years: in 1509 he wrote that

for twelve years now I have gone about all over Italy, leading a miserable life; I have borne every kind of humiliation, suffered every kind of hardship, worn myself to the bone . . . solely to help my family

The choice of 1497 as the year his troubles began is repeated in a letter to his father of 1512:

<pb n="43">

I live meanly . . . with the greatest toil and a thousand worries. It has now been about fifteen years since I have had a happy hour; I have done everything to help you, and you have never recognized it or believed it. God pardon us all.

We have only the <anchor id="1">Bacchus to show for the block Michelangelo was carving for Riario, for the block he bought and worked for himself, and for the commission from Piero de' Medici. There are records of a standing Cupid (perhaps an Apollo) with arrows and quiver, also done for his friend Jacopo Galli. This statue, described as life−size, with a vase at its foot, has disappeared without a trace.

---

2. Michelangelo Sample: Comparison of Automated and Manual Parsing

The tables below show noun phrases, keywords, place names, non−place proper names, dates and related targets (i.e. works) identified in the same sample by the computational linguistic software tools ("Automated" column) and by direct human identification of descriptors ("Manual" column). It also shows summary statistics and calculates "recall" and "precision" for the automated results.

Results are affected, e.g., by lack of specificity as to how noun phrases were to be defined and whether proper names used as modifiers should be consider part of a noun phrase, or as proper names, or both.

" Noun Phrases"		"Keywords"	
Automated	**Manual**	**Automated**	**Manual**
God pardon			
	High Renaissance art		Apollo
Humanistic circle			August
Moses perhaps spring			Bacchus
Roman gentleman	Roman gentleman		Bartholomew
Some of			Cardinal
This statue			Condivi
affairs			David
	alert little satyr		Dionysus−Bacchus
also such men			Galli
ancient cults			Galli's
ancient writers	ancient writers		God
	androgynous antique character		Heemskerck
animal			High
another five ducats			Humanistic
another piece	another piece		Iacopo
	antique sculpture		Italy
antiquity			Jacopo
appetite			Judgement
arrows			July
	arrows and quiver		Last
aspect			Maerten
attributes			Medici
banker			Messer
bearing			Michelangelo
	beneficial influence of antiquity		Michelangelo's
block			Moses
bone			Nevertheless
both in poetry			Phaedrus
brother			Piero
bunch			Quattrocento
	bunch of grapes		Renaissance
century			Riario
	chief view		Riario's
choice			Roman
commission			Rome
commissions			Sadoleto
complaints			Satyrs
	coy little satyr		Silenus
cup			Some
darker things			St
death			Vasari
description			addition
	earliest works	affairs	affairs
eating			alert
eerie figure	eerie figure		allows

every kind			always
every part corresponds			analyze
everything			ancient
expenses			androgynous
	exquisitely graceful	animal	animal
eyes			another
	familiar ancient proportions		antique
family		antiquities	antiquities
	fascinating torsion	antiquity	antiquity
feet			appeal
fifteen years	fifteen years	appetite	appetite
figure		arm	arm
first disconcerting			around
	first masterpiece	arrows	arrows
	first time in modern sculpture		art
five ducats	five ducats		art−poses
flayed skin		aspect	aspect
foot			associated
form			astonished
	form and bearing		attitudes
	free attitudes		attracted
	free−standing statue	attributes	attributes
frenzy			away
	frontal position		back
	full of movement	banker	banker
	further commissions		base
further letter	further letter	bearing	bearing
garland			begs
	garland of vine leaves		believed
genuine unhappiness	genuine unhappiness		beneficial
	god of wine		between
	good judgement and design	block	block
good piece	good piece		body
good understanding	good understanding	bone	bone
grapes			borne
	great personages		both
	greatest toil		bought
grisly orgies			broken
happy hour	happy hour	brother	
hardship			brother's
he		bunch	bunch
head			call
height			carve
heroic works			carved
him			carving
himself		century	century
his			character
	his brother's death		chief

	his own sacred fluid
hooves	
house	
humiliation	
iconographical information	
	ideal nude body
image	
	image of a flayed skin
intimate	
inventor	
it	
its	
ivy	
labor	
	large cup
later years	later years
lawsuit	
left arm	left arm
left hand	
legendary	
	legendary leopardus
letter	
life	
liquor	
little satyr	
looks	
love	
	love of wine
marble	
me	
merry	
miserable life	miserable life
modern antiquities	modern antiquities
money	
mouth	
my	
	my affairs
	my family
	my labor
	my own pleasure
myself	
mystery	
	new spontaneity
	newly−discovered 'appeal and vigor of living flesh
older man	older man
one	
open	

choice	choice
circle	circle
	citing
	collection
	combined
	come
commission	commission
commissions	commissions
complaints	complaints
	composition
corresponds	corresponds
	could
	coy
	crowned
cults	cults
cup	cup
	darker
	de
death	death
	dedicated
	deflect
	degrees
	delights
	depiction
	derived
	described
description	description
	design
	desired
	detail
	detect
	dialogue
	did
	disappeared
disconcerting	disconcerting
	down
	drawing
	drink
ducats	ducats
	early
eating	eating
	eerie
	emphasizes
	encouraged
	enjoys
	enraptured
everything	everything
	evocative

open and eyes			excessively
own and a figure			expense
own pleasure		expenses	expenses
own plight			exquisitely
pagan		eyes	eyes
piece			familiar
	piece of marble	family	family
pleasure			fascinating
	pointed base		father
polished surface			father's
quiver		feet	feet
	raised cup		fifteen
raw flesh		figure	figure
reason			figures
records			filtered
reports			finished
response			flayed
revelry		flesh	flesh
right hand	right hand		fluid
ritual sacrifice			following
sculptors		foot	foot
sculpture		form	form
	sculpture garden		found
sense			free
senses			free−standing
	several points of view	frenzy	frenzy
skin			friend
slightly tipsy young god			frontal
	slow movement	fruit	fruit
soft			full
	standing marble statue		fullness
statue			furnishes
suburban villa			further
symbol			furtively
	symbol of his own plight		garden
teacher		garland	garland
ten palms		gentleman	gentleman
	ten palms in height		genuine
that fruit		god	god
these great personages one			god's
they			gone
think			good
those of people			grace
thousand worries	thousand worries		graceful
tiger		grapes	grapes
	tiger's skin		great
trace			greatest
	traditional art−poses of antiquity		grisly

	transition
troubles	
twelve years	twelve years
	his father
us	
vase	
vine	
we	
wine	
work	
year	
you	

Count	
149	71
Number of Exact Matches: 19	

	had
	half
hand	hand
	happy
hardship	hardship
he	
head	head
height	height
	help
	heroic
him	
himself	
his	
	holds
hooves	hooves
hour	hour
house	house
humiliation	humiliation
	iconographical
	ideal
	illustration
image	image
	imagine
	included
	incorporate
	incorporated
	influence
information	information
	instead
intimate	intimate
	into
inventor	inventor
it	
its	
ivy	ivy
	judgement
kind	kind
labor	labor
	large
	lascivious
lawsuit	lawsuit
	lead
	leading
	learned
	least
	leave
	leaves
legendary	legendary

	leopardus
letter	letter
life	life
	life−size
liquor	liquor
	live
	living
	looking
looks	looks
	loses
love	love
	lovingly made
	made
	make
man	man
marble	marble
	masterpiece
me	
	meanly
	means
	measured
men	men
merry	merry
	mind
	miserable
	modern
money	money
mouth	mouth
	movement
my	
myself	myself
	mysterious
mystery	mystery
	naive
	naked
	natural
	naturalistic
	needed
	neighbor
	never
	new
	newly−discovered
	noble
	none
	novel
	nude
	obviously
of	

	off
	older
one	
	only
open	open
orgies	orgies
	out
	over
	overcome
pagan	pagan
palms	palms
pardon	pardon
	part
	partly
people	people
	personages
piece	piece
	planned
pleasure	pleasure
plight	plight
poetry	poetry
	pointed
	points
	polished
	position
	pressed
	probably
	producing
	promised
	proportions
	pushed
	quite
quiver	quiver
	raised
	rather
	raw
	realize
reason	reason
	recognized
records	records
	referred
	reimbursed
	repeated
reports	reports
	representations
	represented
	reputed
response	response

	returned
revelry	revelry
	ritual
	rolling
	roundness
	sacred
sacrifice	sacrifice
	satisfied
satyr	satyr
sculptors	sculptors
sculpture	sculpture
sense	sense
senses	senses
	set
	several
	show
	shown
	signify
	simply
skin	skin
	slightly
	slow
soft	soft
	solely
	soon
	spirit
	spontaneity
spring	spring
	squinting
	standing
statue	statue
	suburban
	such
	suffered
	support
surface	surface
	survive
	suspect
	suspiciously
	sway
symbol	symbol
	symbolize
	taking
teacher	teacher
	ten
they	
things	things
think	think

	thousand
	threatened
	thrown
tiger	tiger
	tiger's
	time
	tipsily
	tipsy
	toil
	torsion
trace	trace
	traditional
	transition
troubles	troubles
	twelve
	ultimately
understanding	understanding
	undertook
unhappiness	unhappiness
	unusual
us	us
	van
vase	vase
	veiled
	view
	viewer
	vigor
villa	villa
vine	vine
	way
we	
	weathered
wine	wine
work	work
	worked
	working
works	works
	worn
worries	worries
	would
	wreathed
	writer
writers	writers
	writing
	wrote
year	year
years	years
	yet

	young
	accompanying
Count	
133	419
Number of Exact Matches: 120	

"Place Names"	
Automated	**Manual**
Italy	Italy
Roman	
	Rome
Count	
2	2
Number of Exact Matches: 1	

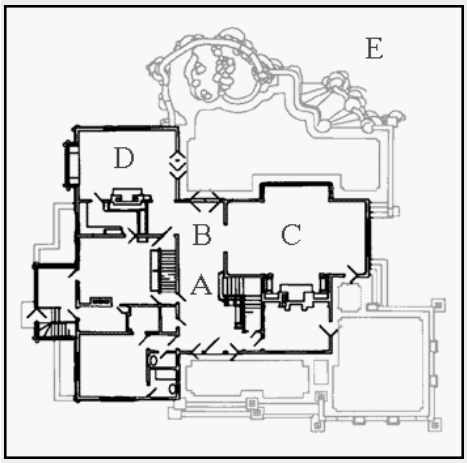
"Non−Place Names"	
Automated	**Manual**
Apollo	Apollo
Bacchus	Bacchus
Cardinal	
Cardinal Riario	
Condivi	Condivi
Cupid	Cupid
David	David
	Dionysus−Bacchus
Galli	
God	God
Humanistic	
Jacopo Galli	Jacopo Galli
Jacopo Sadoleto	Jacopo Sadoleto
Judgement	
	Maerten van Heemskerck
Messer Iacopo Galli	Messer Iacopo Galli
Michelangelo	
	Moses
Phaedrus	Phaedrus
Piero	
	Piero de' Medici
Riario	Riario
	Sadoleto
Satyrs	Satyrs
Silenus	Silenus
St Bartholomew	St Bartholomew
	Vasari
	Michelangelo
father	
Count	

22	21
Number of Exact Matches: 14	

"Dates"	
Automated	**Manual**
1497	1497
1509	1509
1512	1512
	1532−5
Count	
3	4
Number of Exact Matches: 3	

"Related Targets"	
Automated	**Manual**
	#30. Michelangelo, Buonarroti, 1475−1564. Last Judgement
Count	
0	1
Number of Exact Matches: 0	

Appendix C: Greene & Greene Architectural Project Used for CLiMB Exploratory Study



Project 184: ***L.A. Robinson House (Pasadeno, Calif.)*** Residence for L.A. Robinson at Pasadena, Calif. / Greene and Greene, Arch'ts. 1906−[19]11; 70 sheets various media 104.5 x 106.7 cm. (41 1/8 x 42 in.) or smaller.

Text Processed for CLiMB Study

[From: Bosley, Edward. Greene & Greene, Phaidon Press, Inc., 2000]

Chapter 4: Stones of the Arroyo

The <anchor id="184">Robinson dining−room and living−room furniture is where one can best appreciate the significantly more refined work that was now being contributed by the Hall brothers. These pieces should be compared with the Greenes' pieces for <anchor id="214"> Adelaide Tichenor, which were being made at about the time the <anchor id="184"> Robinson house was in the design phase, though the <anchor id="184"> Robinson furniture probably dates to early 1907.27 <anchor id="184">The Robinson dining chair design is derived from early Ming Dynasty furniture: a gentle bow in the crest rail and a "lift" in the bottom stretchers. Corner brackets derive from Japanese construction and were pictured in the book by Edward Morse that Charles bought in late 1903. The crest rail is designed as a flowing line, continuous with the upright members. Other pieces are detailed with traditional mortise−and−tenon joinery, though without direct expression. The sleek simplicity of the chairs is echoed by the two sideboards (the larger designed in 1906 and the smaller in 1910) , which are neither Ming nor Japanese, but neoclassical in mass, with only hints of Chinese influence in the flowing bands of a "lift" motif in relief across the cabinet doors, and Japanese influence in the brackets, similar to those used in temple construction. By contrast, the dining table expresses its construction directly, by way of protruding tenons in the pedestal structure, rivet−like pegs that conceal screws that attach the edge to the top, and visible butterfly joints that join the mahogany slabs of the tabletop. The shape of the top relates to the Japanese tsuba, sword−guard shapes collected by Charles Greene. A radical transformation had taken place since the designs of the relatively foursquare <anchor id="214"> Tichenor furniture. The shapes had become softer, the wood more subtly grained, and the construction far superior. The working relationship between the Greenes and the Halls had sparked a higher level of design and manufacturing. The designs demanded fine craft, but the new availability of craft expertise probably suggested to the Greenes designs they would not have attempted otherwise.

The design and construction of leaded art glass for the <anchor id="184"> Robinson house was further evidence of a dramatic evolution in the Greenes' work. It was at about this time that the Greenes began to contract with Emil Lange, a German−born art−glass craftsman who had relocated to Los Angeles from Burlington, Iowa, in 1904, following a stormy separation from his wife. Lange's Iowa business was mainly in ecclesiastical windows, for which he was responsible for design as well as manufacturing. In Los Angeles he went into business with Harry Sturdy, and their firm, Sturdy−Lange, was soon known for superior work in art glass. The decorative glass in the <anchor id="184"> Robinson house appears similar to the work of Emil Lange and is most stunningly illustrated by the adjustable−height dining−room chandelier.28 The design is a detail of a cherry tree, its fruit hanging in pairs from spreading branches. Unlike Charles Greenes' earlier designs in leaded glass, which were either broadly scenic or depicted highly focused details abstracted from nature (a single flower or leaf), the <anchor id="184"> Robinson designs for the chandelier and the entry doors are like middle−field snapshots of identifiable natural elements, such as most of a tree or a length of vine. This was an important choice by Charles Greene that distinguished his leaded art−glass designs from those by other architects, and may have been made possible by Lange's expertise. Charles did not completely abandon the more tightly focused glass compositions, however, and the stylized feather design in windows for the <anchor id="184"> Robinson den and upper−level hall are close to Prairie School designs in their abstraction and detail.

Comparison of Automated and Manual Parsing

The tables below show noun phrases, keywords, place names, non−place proper names, dates and related targets (i.e. works) identified in the same sample by the computational linguistic software tools ("Automated" column) and by direct human identification of descriptors ("Manual" column). It also shows summary statistics and calculates "recall" and "precision" for the automated results.

Results are affected, e.g., by lack of specificity as to how noun phrases were to be defined and whether proper names used as modifiers should be consider part of a noun phrase, or as proper names, or both.

"Noun Phrases"		**"Keywords"**	
Automated	**Manual**	**Automated**	**Manual**
Chinese influence	Chinese influence	abstraction	
Corner brackets		architects	
German−born art−glass craftsman	German−born art−glass craftsman	availability	
Greenes designs		book	
Hall brothers		bow	bow
Iowa business			bracket
Japanese construction	Japanese construction	brackets	brackets
Japanese influence			branch
Japanese tsuba	Japanese tsuba	branches	branches
	Ming Dynasty furniture	brother	
Other pieces		burlington	
Robinson den		business	
Robinson designs			chair
Robinson dining chair design		chairs	chairs
Robinson furniture		chandelier	chandelier
Robinson house		choice	
These pieces		compositions	compositions
abstraction		construction	construction
			craft

adjustable−height dining−room chandelier	adjustable−height dining−room chandelier
art glass	art glass
bands	
book	
bottom stretchers	
brackets	
branches	
	broadly scenic highly focused details
broadly scenic or details	
business	
cabinet doors	cabinet doors
chairs	
chandelier	
cherry tree	
close	
construction	
contrast	
	corner brackets
craft expertise	craft expertise
crest rail	crest rail
decorative glass	decorative glass
	den and upper−level hall
design	
	design and construction
	design and manufacturing
design phase	design phase
designs	
detail	
	detail of a cherry tree
	dining chair design
dining table	dining table
	dining−room [furniture
dining−room and living−room furniture	
direct expression	
dramatic evolution	dramatic evolution
earlier designs	
early Ming Dynasty furniture	
ecclesiastical windows	ecclesiastical windows
edge	
entry doors	entry doors
expertise	
feather design	
fine craft	fine craft
firm	
	flowing line

	craftsman
	den
design	
	detail
details	details
	dining
	door
doors	doors
	element
elements	elements
evidence	
evolution	evolution
expertise	expertise
expression	
flower	flower
fruit	
furniture	furniture
glass	glass
greenes	
hall	hall
hints	
influence	influence
iowa	
japanese	
joinery	joinery
	joint
joints	joints
lange	
	leaf
level	
lift	
line	line
manufacturing	manufacturing
mass	mass
	member
members	members
morse	
motif	motif
nature	
pairs	
pieces	
	rail
rails	
relationship	
relief	relief
school	
screws	
separation	

	from spreading branches
fruit	
further evidence	
gentle bow	gentle bow
glass compositions	
he	
higher level	
his	
identifiable natural elements	identifiable natural elements
important choice	
its	
	leaded art glass
leaded art−glass designs	leaded art−glass designs
leaded glass	leaded glass
leaf	
length	
	length of vine
lift	
	lift motif
line	
	living−room furniture
mahogany slabs	mahogany slabs
manufacturing	
mass	
middle−field snapshots	middle−field snapshots
motif	
nature	
	neoclassical in mass
new availability	
one	
only hints	
pairs	
pedestal structure	pedestal structure
pieces	
place	
	protruding tenons
radical transformation	radical transformation
relationship	
relatively foursquare Tichenor furniture	
relief	
rivet−like	
	rivet−like pets
screws	
shape	
significantly more refined work	
single flower	
	single flower or leaf chandelier

	shape
shapes	shapes
sideboard	sideboard
simplicity	simplicity
	slab
slabs	slabs
	snapshot
snapshots	snapshots
	stretcher
stretchers	stretchers
structure	
sturdy	
table	
tabletop	tabletop
tenons	
timber	
time	
transformation	transformation
tree	tree
	tsub
tsuba	
	vine
ways	
willett	
	window
windows	windows
work	
Count	
73	56
Number of Exact Matches: 33	

sleek simplicity	sleek simplicity
stormy separation	
	stylized feather design
superior work	
sword−guard shapes	sword−guard shapes
tabletop	tabletop
temple construction	temple construction
tenons	
their	
they	
those by other architects	
	tightly focused glass compositions
time	
top	
traditional mortise−and−tenon joinery	traditional mortise−and−tenon joinery
tree	
two sideboards	
upper−level hall	
upright members	upright members
vine	
visible butterfly joints	visible butterfly joints
way	
wife	
windows	
wood	
work	
Count	
115	52
Number of Exact Matches: 31	

"Place Names"	
Automated	**Manual**
Burlington	Burlington
Iowa	Iowa
Los Angeles	Los Angeles
Count	
3	3
Number of Exact Matches: 3	

"Non−Place Names"	
Automated	**Manual**
Adelaide Tichenor	Adelaide Tichenor
Charles	Charles
Charles Greene	Charles Greene
Charles Greenes	
Edward Morse	Edward Morse
Emil Lange	Emil Lange

Greenes	Greenes
Hall	
	Hall brothers
Halls	
Harry Sturdy	Harry Sturdy
It	
Japanese	
Lange	
	Ming
Ming Dynasty	Ming Dynasty
Prairie School	Prairie School
Robinson	Robinson
	Sturdy−Lange
Tichenor	
	the Halls
Count	
17	14
Number of Exact Matches: 10	

"Dates"	
Automated	**Manual**
1904	1904
1906	1906
1910	1910
early 1907	early 1907
late 1903	late 1903
Count	
5	5
Number of Exact Matches: 5	

"Related Targets"	
Automated	**Manual**
184	184
214	214
Count	
2	2
Number of Exact Matches: 2	

Appendix D: Background Material on the Use of Computational Linguistic Techniques for Text Analysis

The purpose of this Appendix is to provide a brief overview of the field of text processing and computational linguistics in order to provide some context for the proposal. The research topics include such areas as the analysis of authorship, stylistic text analysis, and dictionary creation. A brief description of these methods and their applications will help to clarify the background of this proposal and will illustrate the pool from which we will draw our methodology.

To our knowledge, computational text analysis techniques have not yet been applied to the problem of indexing. Computational linguistics is a relatively new field, combining the traditional area of language analysis and linguistic theory with the newer area of computer science. (For further reading, please refer to Klavans 1989.) The resulting hybrid discipline has seen results applied to areas such as the following, many of which are relevant to this proposal.

- ***Language Identification*** – Language analysis techniques have been used to solve the problem of looking at a page of text and figuring out what language it is written in. Clues such as type of alphabet, distribution of characters, and even sentence length have been used. For example, a sentence in English when translated into French tends to become longer. The same document in French can be nearly 30% longer than its English counterpart.

- ***The Analysis of Authorship*** (and its darker side of the analysis of plagiarism) – The most common techniques used for authorship attribution involve statistical measures applied to words, phrases, and their frequencies of use. Added to these statistical methods are techniques that permit a view into the more subtle aspects of syntax, such as clause structures favored by a particular author. Results from computational linguistic authorship studies have been applied to problems as well studied as Shakespeare authorship.

- ***Disputed Authorship*** – This is a major area in forensic linguistics, where computational linguistic methods are used to compare word usage in two documents of similar length where one author is known and the other unknown, generating a probability that the two documents were produced by the same individual.

- ***Stylistic Text Analysis*** – The study of stylistics has used many of the same computational linguistic techniques as authorship analysis, but with the goal of categorizing documents in a collection by style. This requires a theory of style and genre, such as news, fiction, and even sub−styles such as historical vs. biographical fiction.

- ***Usage of Words and Phrases for Dictionary Creation*** – Text analysis has been used for the creation of large collocational dictionaries that have been incorporated in published dictionaries such as the groundbreaking CoBuild Dictionary of English. For the first time, when computer disk space became more available in the 1980s, a large 100 million−word corpus was collected and analyzed. Studies of English language usage were used to collect realistic examples of words and phrases as they are actually used. Similar techniques are now used to study language variation and differences between dialects such as British, Indian, and American English.

- ***The Creation of Bilingual Dictionary Data*** – Translation is a time−honored discipline, but increasingly, especially for publication on the Internet, many documents are created on−line in more than one language at a time. For example, all Swiss government documents must be in English, Italian, German and French. By identifying the way words and phrases are

translated, large multilingual dictionaries are now being created automatically, instead of relying solely on the individual translator. Such tools add to the translators' resources, helping them to create more accurate translations.

Appendix E: Staffing Details

This Appendix presents a detailed explanation of the function of each of the team members presented in Section 13 of the proposal.

Staffing will include:

1. **Computational Linguistics Specialist/Project Manager (Klavans −−20%)**

   Coordination responsibilities include:

   a. management of project team
   b. establishing and tracking workflow for team members
   c. building the infrastructure for project
   d. integrating with external projects
   e. overseeing establishment and meeting of target deadlines for components
   f. ensuring that project results are published and presented nationally and internationally

   Research responsibilities include:

   g. direct supervision of computational linguistic research
   h. overseeing incorporation results into platform for assessment by scholars and user groups
   i. publishing and presenting research results in major academic venues

2. **Digital Collections/Art Research Librarian (100%)**

   Chief Responsibilities

   a. Perform analyses of image cataloging, indexing and retrieval techniques for innovative study of computer−assisted metadata generation
   b. Participate in selecting image collections and corresponding scholarly texts for test system;
   c. Advise in selection of image storage and retrieval system to use as project demonstration system
   d. Help design and conduct end−user assessment & use study of demonstration system
   e. Facilitate communication among project partners, advisory board, library staff and scholars and researchers

   Qualifications

   f. Master's degree in library science and 3−4 years experience as art librarian, curator or cataloger; advanced degree in Art History or comparable field desirable
   g. Experience with image cataloging, indexing and retrieval systems
   h. Experience in building digital image collections
   i. Experience in working with scholars and other end users in research involving image collections

3. **Programmer/Analyst (50%)**

The programmer will be responsible for incorporating the results of output from CLiMB tools into existing platforms for assessment. This will include:

a. evaluating a set of potential platforms and tools as used by image collection specialists, with particular attention and coodination with ArtStor developers
b. incrementally incorporating the output of CLiMB tools into chosen platforms
c. designing, with team members, the methods of querying, displaying and browsing of larger metadata
d. working with other team members to build the evaluation test set, which will be used as a standard for measuring progress
e. given this manually built standard, building automatic tools to collect assessment data (e.g. scripts for collecting user input and feedback, alternative displays for user evaluation)
f. ensuring that all tools are publicly available and supported

4. **Computational Linguistics Graduate Student (12 month Ph. D. student)**

The computational linguistics graduate student will be responsible for customizing existing tools to refine output for the descriptive metadata application and developing new capability when needed. This will include:

a. selection of a set of existing tools to test functionality
b. development of evaluation techniques to measure effectiveness
c. building of software to automatically measure improvements
d. incorporation of input from users on stoplists and lists for boosting work and phrase rankings
e. packaging of tools for export and use to other sites

Graduate students in the Department of Computer Science at Columbia University spend 20 hours per week on course and class work and 20 hours on research. The student will be directly supervised by Klavans, and will be a member of the Natural Language Processing group.

5. **Digital Projects Implementation Librarian (Davis − 10%)**

a. Coordinate the selection and implementation of demonstration image storage and retrieval system(s)
b. Advise in the area of metadata standards, structures and interoperability
c. Act as project liaison to Library Systems Office, Academic Information Systems and Bibliographic Control Dept.

6. **Project Assistant (50%)**

The function of the project assistant is to collect information from team members at the end of each semester, i.e. December and May, and in August in order to collate project reports. This person will also be responsible for interfacing with other project participants to make sure interchange of data is smooth. As an interdisciplinary project, such coordination helps to document progress and ensure smooth collaboration. It also contributes to effective communication of progress to outside collaborators, and thus to wide dissemination of results. The assistant will maintain an active web site as part of this function.

We will request travel for an advisory board to meet on an annual basis, which will be part of our plan for dissemination of results.

Appendix F: Testbed Collections

Below are descriptions of our preliminary selection of testbed collections and texts to be used for CLIMB. As discussed elswhere in the proposal, there may need to be additions or changes during the course of the project, but we believe these represent an adequate critical mass of material for testing.

1. Greene & Greene Virtual Archive

Columbia's portion to the Greene & Greene Virtual Archive Project consists of about 4700 digitized architectural drawings and around 400 photographs. The composite database from Columbia, Berkeley, University of Southern California and the Huntington will be about twice that. The target completion date is mid−2002. See also Appendix C for a fuller description of the Greene & Greene project and results of preliminary testing.

In the original CLIMB proposal we processed a sample from the recent book "Greene & Greene" by Edward Bosley (London : Phaidon, 2000). In addition to this work, we would be able to use these additional texts Columbia's Collections:

- **Makinson, Randell L., 1932−** Greene & Greene. Salt Lake City : Peregrine Smith, c1977−1979. 2v.
- **Current, William R.** Greene & Greene; architects in the residential style. Fort Worth <Tex.> Amon Carter Museum of Western Art <1974>. 128p.
- **Smith, Bruce, 1950−** Greene & Greene : masterworks. San Francisco : Chronicle Books, 1998. 240p.
- **Makinson, Randell L., 1932−** Greene & Greene : the passion and the legacy. Salt Lake City : Gibbs Smith, c1998. 231p.
- plus a variety of smaller descriptive pamphlets and exhibition catalogs

Because the Greene & Greene Virtual Archive will itself include standard descriptive metadata, it will provide an excellent basis for comparing CLIMB generated metadata with finding aid−type description; and also testing CLIMB metadata as an enhancement to existing descriptions.

Angela Giral, Director of Columbia's Avery Library will coordinate scholarly input for this part of CLIMB.

---

2. American Institute of Indian Studies (AIIS), Center for Art & Archaeology, Photo Archive

Part of the Digital South Asia Library, the AIIS collection from the Center for Art and Archaeology in Gurgaon, Haryana, India, has over 125,000 photographs. These images fall into the broad categories of architecture, sculpture, terracotta, painting and numismatics. We would target architecture, sculpture and terracotta for this project.

David Magier, Columbia's Director of Area Studies and a primary coordinator of the Digital South Asia Library, will coordinate scholarly input for this part of CLIMB. Columbia's Avery Library has a rich collection of monographs and descriptive works on Indian and South Asian art and architecture which alone could supply a sufficient number of texts for use by CLIMB.

Below is a preliminary, very partial listing of specific subsets of the AIIS collection that would be used for CLIMB, along with selected scholarly and descriptive texts from the Avery collection.

i. **Sun Temple, Konarak** (Puri, Orissa, India): 111 photos

Relevant texts in Avery Library:

a. **Behera, Karuna Sagar.** Konarak : the heritage of mankind. New Delhi : Aryan Books International, 1996. 2v.
b. **Mitra, Debala, 1925−** Konarak. New Delhi : Archaeological Survey of India, 1976. 124p.
c. **Boner, Alice.** New light on the Sun Temple of Konarka; four unpublished manuscripts relating to construction history and ritual of this temple, 1972. 238p.

etc.

ii. **Halebid, Hassan** (Karnataka, India): 58 photos

Relevant texts in Avery Library.

a. **Evans, Kirsti.** Epic narratives in the Hoysala temples : the Ramayana, Mahabharata, and Bhagavata Purana in Halebid, Belur, and Amrtapura. Leiden ; New York : E.J. Brill, 1997. 286p
b. **Maity, Sachindra Kumar.** Masterpieces of Hoysala art : Halebid, Belur, Somnathpur. Bombay : Taraporevala, 1978. 52p.

iii. **Khajuraho, Chhatarpur** (Madhya Pradesh, India): 74 photos

Relevant texts in Avery Library.

a. **Desai, Devangana.** Khajuraho. New Delhi : Oxford University Press, 2001. 107p.
b. **Stierlin, Henri.** Hindu India : from Khajuraho to the temple city of Madurai. Koln ; New York : Taschen, c1998. 237p.
c. **Suresh, K. M., 1952−** Saivite sculptures of Khajuraho. Delhi : Bharatiya Kal Prakashan, 1998. 147p.
d. **Khanna, Ashok.** Rhythm in Khajuraho. Delhi : South Asia Publications, 1997. 160p.
e. **Desai, Devangana, 1937−** The religious imagery of Khajuraho. Mumbai : Franco−Indian Research, c1996 269p.
f. Khajuraho in perspective : Proceedings of the U.G.C. National Seminar on "Art of Khajuraho". Bhopal : Commissioner, Archaeology and Museums, Madhya Pradesh, 199. 238p.
g. **Majupuria, Trilok Chandra.** Glories of Khajuraho : a description of the unique art and architecture of some of the magnificent temples of medieval India. Lashkar, Gwalior, India : M. Gupta, 1990. 344p.
h. **Krishna Deva, 1914−** Temples of Khajuraho. New Delhi : Archaeological Survey of India, c1990. 521p.
i. **Lal, Kanwar.** Apsaras of Khajuraho. Delhi, Asia Press <1966>. 34p.
j. **Agarwal, Urmila.** Khajuraho sculptures and their significance. Delhi, S. Chand, 1964. 220p.

iv. **Mahabalipuram, Chingleput** (Tamilnadu, India): 144 photos

Relevant texts in Avery Library:

a. Descriptive and historical papers relating to the seven Pagodas on the Coromandel Coast. New Delhi : Asian Educational Services, 1984. 242p.
b. **Sivaramamurti, C**. Mahabalipuram. New Delhi : Archaeological Survey of India, 1978. 35p.
c. **Srinivasan, K. R., 1910−** The Dharmaraja ratha & its sculptures, Mahabalipuram. New Delhi : Abhinav Publications, 1975. 112p.
d. **Lockwood, Michael, 1933−** Mahabalipuram studies. Madras : Christian Literature Society, 1974. 111p.

v. **Sanchi, Raisen** (Madhya Pradesh, India): 14 photos

Relevant texts in Avery Library:

a. Unseen presence : the Buddha and Sanchi. Mumbai : Marg Publications, 1996. 134p.
b. **Rao, Manjushri.** Sanchi sculptures, an aesthetic and cultural study. New Delhi : Akay Book Corp. ; Delhi : Distributor, Vidyanidhi Prakashan, 1994. 221p.
c. **Srivastava, A. L.** Life in Sanchi sculpture. New Delhi : Abhinav Publications, 1983. 163p.
d. **Marshall, John Hubert, 1876−1958.** The monuments of Sanchi. Delhi : Swati Publications, 1982 (rep. of 1940 ed.). 3v.
e. **Cunningham, Alexander, 1814−1893.** The Bhilsa topes; or, Buddhist monuments of Central India; comprising a brief historical sketch of the rise, progress, and decline of Buddhism, with an account of the opening and examination of the various groups of topes around Bhilsa. Varanasi, Indological Book House, 1966. 236p.
f. **Maisey, Fredrick Charles.** Sanchi and its remains : a full description of the ancient buildings, sculptures, and inscriptions at Sanchi, near Bhilsa, in Central India, with remarks on the evidence they supply as to the comparatively modern date of the Buddhism of Gotama, or Sakya Muni. London, K. Paul, Trench, Trubner, 1892. 142p.

Other relevant general texts from Avery Library, to be evaluated for use in CLiMB project.

- **Ramachandra Rao, Saligrama Krishna, 1926−** Art and architecture of Indian temples. Bangalore : Kalpatharu Research Academy : UBS Publishers' Distributors <distributor>, 1993− 3 v

- **Ramachandra Rao, Saligrama Krishna, 1926−** Indian temple traditions. Bangalore : Kalpatharu Research Academy, c1997. 346 p.
.
- Agama−kosha = Agama encyclopaedia. Bangalore : Kalpatharu Research Academy : Sole distributors, T.N. Krishnaiah Setty & Sons, 1989−1994. 12 v.

- **Dei, Shashipriya, 1943−** Development of temple architecture in India : with reference to Orissa in the golden age. Calcutta : Punthi−Pustak, 1998. 116p.

- **Stierlin, Henri**. Hindu India : from Khajuraho to the temple city of Madurai. Koln ; New York : Taschen, c1998. 237p.

- **Grover, Satish, 1940−** The architecture of India : Buddhist and Hindu. Ghaziabad, India : Vikas Pub. House, 1980. 240p.

- **Moorthy, K. K**. Sarvam Sakti mayam : a mini compendium of 300 Sakti temples. Tirupathi : Message Publications, 1997. 324p.

---

3. Chinese Paper Gods

Chinese paper gods are woodblock prints representing a great variety of deities, from diverse pantheons, drawing from Taoism, Buddhism, Confucianism, and animism; deified historical and legendary characters are included as well. Even characters from literature and opera have been adopted as deities in some cases. Prints have been produced and used for many centuries in many parts of China. Although originally they served mainly a religious, ritual function, today they are of particular interest and importance both as art and as sources of information in a variety of fields − cultural history, religious history, the history of print making, paper making, and woodblock carving, and the literary tradition. The prints are not just of interest to scholars and specialists: their visual interest and layers of symbolic meaning make them an excellent tool for introducing audiences of all ages to Chinese civilization.

The Anne S. Goodrich Collection in the C. V. Starr East Asian Library was collected in 1931, when Anne Goodrich bought out the entire stock − over one hundred prints − of a single local print shop in Beijing, and began her study of the layers of meaning in the prints. Sixty years later she published what is now a standard reference on the subject and based on this collection:

> **Goodrich, Anne S.** *Peking paper gods: a look at home worship.* (Monumenta serica monograph series ; 23) Nettetal: Steyler Verlag, 1991.

Since they were produced to be used and discarded, on cheap paper, frequently burned at New Year's, there are few holder print extant.

Columbia is in the process of conserving approximately 110 images from the Goodrich collection, using techniques specifically developed for these prints by a paper conservator. The images are then being digitized to provide broader access to the materials and reduce physical handling.

Amy Heinrich, Director of Columbia's Starr East Asian Library, will coordinate scholarly input for this part of CLIMB.

---

4. Images from the Academic Image Cooperative Database

The Academic Image Cooperative database was initiated through the Digital Library Federation in January 1999, with funding from Mellon, to enable a new kind of community building amongst art historians and visual resources professionals by facilitating the sharing and exchange of art history images and data over Web. As described in the initial AIC document (www.diglib.org) in its initial inception it aspired to be:

- a shared cataloging utility for the visual resource profession, from which image catalogers may freely derive cataloging records and to which image catalogers may also contribute such records readily;
- a shared image library created for and by the art history and visual resource community unconstrained by copyright restrictions; and
- an affordable database, if not indeed made available free of charge, for educational purpose

The over 2000 images were selected for their wide use in education and thus are of particular interest to the CLiMB group. What is of use to CLiMB is the fact that the very selection process required that each image be described in an overlap concordance of at least two of the standard survey texts. This very fact ensures that there is text associated with the selected images. The concordance database indicating which images occur in which textbooks can be used by CLiMB to identify both high−use images as well as related text. The metadata for the AIC collection is especially challenging since the metadata is sparse at present. Thus, the results of the CliMB approach could be highly valuable for increased access.

---

5. Other possible online image collections with associated texts

Additional sites for image collections that we might explore include:

1. Illinois: http://images.library.uiuc.edu/projects/

2. Minnesota: http://digital.lib.umn.edu/IMAGES/

3. Michigan: http://images.umdl.umich.edu/b/borobudur/

   These are similar in some ways to the Dunhuang caves project, with images of a Javanese temple. What might make this suitable for CLiMB is that the images seem to be highly structured, possibly in a way that would link to text. Columbia University Libraries has 27 books on this temple.