# Relation between Agreement Measures on Human Labeling and Machine Learning Performance: Results from an Art History Image Indexing Domain

Rebecca J. Passonneau[1], Tae Yano[2], Tom Lippincott[1], and Judith Klavans[3]

[1] Columbia University
(becky|tom)@cs.columbia.edu

[2] Carnegie Mellon University
tyyano@gmail.com

[3] University of Maryland
jklavans@umd.edu

**Word count: 1,993**

## 1 Introduction

We describe a series of studies aimed at identifying specifications for marking up textual input for an image indexer's toolkit. Given an image and a text extract that describes an art or architectural work depicted in an image, our goal is to identify the semantic function served by a span of text with respect to image description. We illustrate this below in Figure 1.

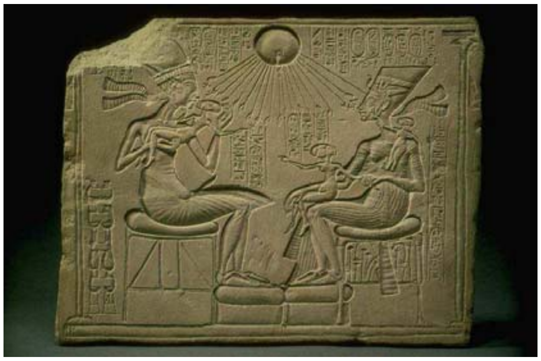
The domain of digital images and texts we focus on parallels the ARTstor *Art HistorySurvey Collection (AHSC),* a Mellon funded collection of 4,000 images. The AHSC is based on thirteen standard art history survey texts, thus there is a strong correlation between the images and these texts. The AHSC images all have tombstone metadata (e.g., the name of the work, the artist, date, the location of the work), but few have subject matter metadata. We are currently using two of the texts from the AHSC concordance that we scanned and encoded in TEI-Lite (http://www.teic.org/Lite/teiu5 split en.html).

In consultation with domain experts, we developed a set of seven functional semantic categories to apply to paragraphs or sentences associated with specific images. Our categories are derived from what we observe in the texts, but have a loose correspondence with categories of information discussed in the image indexing literature [3, 7, 2]. Figure 1 in the next section illustrates three of our seven categories that we have done machine learning for.

Many studies of machine learning use annotated data, and report the levels of interannotator agreement, but do not directly address the question of how much agreement is enough for learnability. For example, in both [4] and [13], as in the present study, human labelers classify spans of text (sentences or sequences of sentences), and machine learners are developed to perform the same classification. Here we report on a series of pilot studies where we developed functional semantic categories to label art history survey texts, to measure interannotator agreement under a variety of annotation task constraints, and to evaluate machine learning performance. We have relatively low agreement overall, but good performance on the categories we have sufficient data for.

Our goal with respect to the image indexer’s toolkit is to find a set of one or more labels that is useful to image indexers, and that an automatic classifier can apply with high reliability to art history survey texts. Our research goals emerge from the observation that for many semantic tasks, there is relatively low interannotator agreement (see [1]). First, we wanted to understand why previous investigators have found such a wide range of interannotator agreement on similar tasks [5,9]. To address this, our pilot annotation studies included four experiments where we varied annotation parameters. Our second research goal is to investigate the relation between agreement and learnability and the nature of our annotation schema.

**Figure 1.** Simplified illustration of semantic classification of text extracts



| | |
|---|---|
| **Historical Context** | Of the great projects built by Akhenaten hardly anything remains . . . Through his choice of masters, he fostered a new style. |
| **Implementation** | Known as the Amarna style, it can be seen at its best in it can be seen at its best in |
| **Image Content** | a sunk relief portrait of Akhenaten and his family. The intimate domestic scene suggests |
| **Historical Context** | that the relief was meant to serve as a shrine in a private household. |

**Enhanced XML representation:**

```
<p><semcat type=”historical”>Of the great
projects built by Akhenaten hardly anything remains</semcat>. .
. . <semcat type=”historical”>Through his choice of masters, he
fostered a new style.</semcat> <semcat type=”implementation”>
Known as the Amarna style, it can be seen at its best in a sunk
relief portrait of Akhenaten and his family.</semcat> <semcat
type=”historical”>The intimate domestic scene suggests that the
relief was  meant to serve as a shrine in a private
household.</semcat> . . ..</p>
```

## 2 Brief Example

Within a paragraph about a given image, the descriptive information can be categorized into distinct types depending on the semantic function of the text with respect to the work depicted in the image. Figure 1 illustrates text from the first part of a few paragraphs associated with an image of a relief portrait of Akhenaten and his family. The image here is taken from the ARTstor Art Images for College Teaching collection (AICT): http://www.arthist.umn.edu/aict/html/ancient/EN/EN006.html. The text fragment is from one of the texts in the concordance to the ARTstor Art History Survey Collection. It has been separated into labeled text spans exemplifying the three categories we have performed learning experiments on. As illustrated, a single sentence can have subparts with distinct semantic functions. The sample of a provisional xml representation shows a sentence-level assignment because we will not attempt to find subspans within sentences.

## 3 Human Labeling

We conducted four pilot studies on the labeling where we varied the number of labels that could be assigned to a single item (one, two or unrestricted), the size of the text fragment being labeled (paragraph or sentence), the number of annotators (two to seven), and the type of training for annotators (none, finished examples presented to trainees, true training examples with feedback). During the pilot studies, we developed a labeling interface, then re-implemented the interface for our ongoing, large scale data collection effort (see [8]).

To measure interannotator agreement, we use Krippendorff's *Alpha* [6] along with a set-based distance measure [10] to allow partial credit when the set of labels chosen by one annotator overlaps another's set. Our measures of interannotator agreement varied widely. These results are consistent with previous literature on interannotator consistency in the library and image cataloging domains [5,9], but in contrast to previous work, we can relate deltas in the amount of agreement to specific causes.

Annotation efforts typically aim for agreement measures above a threshold of 0.67, as suggested by Klaus Krippendorff [6]. We have previously argued that because agreement coefficients do not have a known probability distribution, and because they are applied to many kinds of data from many disciplines (see [1]), there is no single ideal threshold for all cases [10, 11]. Instead, we suggest that it is an empirical question that can be investigated in many ways, for example relating measures of tasks in which the annotations are used to the observed agreement levels. A related point has been made in [12], where Riedsma and Carletta report on simulations of learnability from data with different levels of agreement. They present evidence that performance of machine learners does not correlate directly with agreement levels.

Our results on the human labeling task indicate that reliability improves if annotators can select multiple labels, which is consistent with our previous results on a lexical semantic annotation task [11]. We also find that labeling can be done more consistently by senior domain experts, that sentence level labeling can be done more consistently than paragraph level labeling, and that the consistency of the labeling depends heavily on the image/text pair under consideration. In the first pilot dataset (10 images, 24 paragraphs, 159 sentences) where we collected both sentence and paragraph labeling, the overall interannotator agreement among seven annotators was 0.24 for paragraphs and 0.30 for sentences. In the first dataset from our large scale annotation effort (25 images, 45 paragraphs, 313 sentences) annotated by five individuals including two senior domain experts, overall agreement was 0.40 for paragraphs and 0.47 for sentences. When we computed agreement for all 2 to 5 combinations of annotators, the highs (for the two senior experts) were 0.56 (paragraphs) and 0.55 (sentences). Agreement among five annotators broken down by the image being described ranged from 0.70 to 0.16.

## 4 Machine Learning

Using data from our pilot studies of human labeling, augmented by an additional set of images labeled by one of the co-authors, we investigated the learnability of three categories: Image Content, Historical Context and Implementation. There were insufficient examples from the other categories. All learning was done using WEKA.

We created three types of feature sets. Set A consisted of word features selected on the basis of significant chi-square tests. Set B consisted of hand-picked features in approximately half a dozen groups, such as words and phrases characteristic of the art history domain (e.g., *masterpiece*), and words and phrases referring to parts of the human body. Set C consisted of the union of Sets A and B. Among three types of learning algorithms we tested, naive Bayes, SVM and tree-based classifiers, naive Bayes performed best overall. On ten-fold cross-validation, the highest classification accuracy on Image Content relied on feature set C, and achieved 83% accuracy, compared with 63% for Historical Context and 53% for Implementation. The highest accuracy for Historical Context used feature set A: 70%. Using a random forest classifier for the Implementation class, we achieved an accuracy of 80%.

The most frequent label combination for both paragraphs and sentences was the single label Image Content out of 47 distinct combinations of labels for sentences, and 38 combinations for paragraphs. This partly accounts for the relatively high accuracy of learning for the Image Content classifier.

## 5 Conclusion

Because we are conducting interannotator agreement studies in tandem with machine learning, we can investigate the relationship between the two. We argued (as in [10] that this is an empirical question, given current knowledge. Our results bear out the simulation study presented in [12] that good learning performance can occur when agreement is less than the 0.67 threshold proposed by Krippendorff.  This does not mean that good learning performance never requires higher levels of agreement (see [12]).  Instead, it shows that in richly annotated datasets such as this one, where we have attempted to develop a set of fully covering categories, interannotator agreement and learnablity interact with the distributions of various categories in the labeled data. We discuss this issue in greater length in the full paper.

## References

1. R. Artstein and M. Poesio. Kappa3 = alpha (or beta). Technical Report NLE Technote 2005-01, University of Essex, Essex, 2005.
2. M. Baca. *Practical Issues in Applying Metadata Schemas and Controlled Vocabularies to Cultural Heritage Information*. The Haworth Press, Inc., 2003. Available through Library Literature.
3. H. Chen. An analysis of image queries in the field of art history. *Journal of the American*

*Society for Information Science and Technology*, pages 260–273, 2001.
4. B. Hachey and C. Grover. Sentence classification experiments for legal text summarisation. In *Proceedings of the 17th Annual Conference on Legal Knowledge and Information Systems (Jurix)*, 2004.
5. A. Giral and A. Taylor. Indexing overlap and consistency between the Avery Index to Architectural Periodicals and the Architectural Periodicals Index. Library Resources and Technical Services 37(1):19-44, 1993.
6. K. Krippendorff. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills, CA, 1980.
7. S. S. Layne. Some issues in the indexing of images. Journal of the American Society for Information Science, pages 583–8, 1994.
8. T. Lippincott, T. Yano and R. Passonneau. Submitted. ANTArt: An Extensible Framework for CollectingAnnotations on Texts that Describe Art Images. Submitted to LREC 2008.
9. K. Markey. Interindexer consistency tests: a literature review and report of a test of consistency in indexing visual materials. Library and Information Science Research, pages 155–177, 1984.
10. R. Passonneau. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC), 2006.
11. R. Passonneau, N. Habash and O. Rambow. Inter-annotatator agreement on a multilingual semantic annotation task. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC), 2006.
12. D. Riedsma and J. Carletta. Reliability measurement: there’s no safe limit To appear in Computational Linguistics.
13. S. Teufel and M. Moens. Summarising scientific articles – experiments with relevance and rhetorical status. Computational Linguistics, pages 409–445, 2002.