An empirically-based spatial segmentation and coreference annotation scheme for comics

Lauren Edlin
L.Edlin@qmul.ac.uk
School of Electronic Engineering and Computer Science,
Queen Mary University of London
London, United Kingdom

Joshua Reiss
Joshua.Reiss@qmul.ac.uk
School of Electronic Engineering and Computer Science,
Queen Mary University of London
London, United Kingdom

ABSTRACT

There is recent work in applying concepts from discourse analysis to comics in order to formalise visual content across comics sequences, however much of this work lacks empirical verification. We address this gap by assessing inter-annotator agreement on a preliminary annotation scheme for segmenting areas of comic pages, and assigning referents or classifications to these segmentations. A browser-based annotation tool and inter-annotator agreement measures are introduced. We find high agreement for segmentation tasks, while reference tasks show adequate agreement as well as instructive disagreements showing constraints on reader interpretation.

CCS CONCEPTS

• **Applied computing** → **Arts and humanities**.

KEYWORDS

visual narrative, comics, inter-annotator agreement, comics annotation, discourse analysis, image segmentation, coreference

ACM Reference Format:
Lauren Edlin and Joshua Reiss. 2021. An empirically-based spatial segmentation and coreference annotation scheme for comics. In *The 14th International Symposium on Visual Information Communication and Interaction (VINCI '21), September 6–8, 2021, Potsdam, Germany.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3481549.3481560

1 INTRODUCTION

According to comics practitioner and theorist Scott McCloud, comics are "juxtaposed pictorial and other images in deliberate sequence, intended to convey information and/or produce an aesthetic response in the viewer" [22, p. 9]. The inherent sequential aspect has led researchers to investigate comics as a medium with properties conducive to helping understand general visual communication and cognitive processes (e.g. [8, 14, 19, 22]), and also lends to systematic and corpus-based analyses to test hypotheses regarding these processes. Therefore, applying methods and concepts from computational linguistics (CL) and discourse analysis (DA) to discover patterns in visual information structures across comics appears to be a promising research avenue. While a number of annotated comics corpora are available, their ontologies do not take into account interpretations of readers and lack empirical validation (e.g. [3, 16]). Comics ontologies adapting concepts from DA that do take readers interpretation into account only provide qualitative descriptions of common comics structures, and also lack empirical validation.

This paper addresses this gap of an empirically robust comics ontologies by developing a preliminary comics annotation scheme of page segmentation (panels), location reference, character segmentation and reference, and text section segmentation and classification, and assessing its inter-annotator agreement. The purpose of this work is to assess this initial annotation methodology for efficiency and reliability, including identifying constraints on reader interpretation through annotator disagreement. We investigate the following research questions:

(1) What is an appropriate annotation scheme for efficient marking up of the segmentation and referents/coreferents of comics stories?
(2) Is it possible to achieve high inter-annotator agreement using the scheme, thus giving it empirical validity?
(3) Does a disagreement indicate need for clarification in the scheme, or interpretive ambiguities in the comic narrative?

We approach these questions by developing an annotation scheme that applies concepts from DA by segmenting areas of a comic page [1], and assigning referents/coreferents or other semantic classifications, to these segmentations. We then test the validity of this annotation methodology through an inter-annotator agreement experiment that uses intersection over union (IOU) to assess segmentation agreements and Cohen's κ, a common agreement measure used in CL, to assess coreference agreement. To accomplish the annotation tasks with multiple annotators, a browser-based tool is developed allowing annotators to directly markup a comics page with bounding boxes and assign referents on responsive text-input forms. We find that segmentation with a bounding box tool is reliable for most segmentation tasks, and offer novel thresholds for agreement for each task. We find adequate agreement on many coreference and classification tasks, and find instructive instances of disagreement on character and location reference.

2 RELATED WORK

There has been a recent empirical turn to comics studies [14], including comics annotation. However, most of this work pertains to automatic segmentation tasks or other computer vision applications rather than reader interpretation of comics - see [3, 16] for an overview of these corpora. Bateman et al. have developed an

VINCI '21, September 6–8, 2021, Potsdam, Germany

ACM ISBN 978-1-4503-8647-0/21/09…$15.00
https://doi.org/10.1145/3481549.3481560

extensive classification scheme for comic page layouts [5], which is promising for corpus analyses on visual patterns (e.g. [4]). Annotation schemes involving greater aspects of reader interpretation have been tested and applied to relatively smaller-sized comics corpora (e.g. [13]), however these schemes tend to be very theoretically specific with questionable scalability. Lastly, recent experimentation with crowd-sourcing comics annotations show that computer-based platforms can be used to gather a large number of annotations quickly [29].

In order to explicitly convey reader interpretation into comics analyses, concepts and methods from linguistic discourse analysis (DA) are increasing applied to comics. *Discourse* refers to meaning produced across sentences, rather than focusing on meaning derived from grammatical structures within a sentence. As each sentence reveals new information or previously referenced elements, DAs must account for sentence constituents that refer to the same entities - pronouns, demonstratives, and names, for example - to present an accurate and up-to-date model. Words and phrases that refer to the same entity across sentences are termed *referring expressions*. Two or more referring expressions, or *discourse referents*, that point to the same entity and therefore have the same meaning are *coreferents*. Several researchers have applied aspects of DA to develop formal models of meaning for visual narrative. For instance, [1] and later [20, 21] provide a model of discourse referents constructed out of areas in a picture - that is, the referent is a marking on a specified space in the image, such as the area that depicts a particular character. Another example is [30], who qualitatively describes comic elements and their relations in logical form as abductive inferences that produce meaning.

Much of the work on applying DA concepts to comics remains theoretical, and have assumed ontologies and categorisation of visual elements without robust empirical verification. Coreference assignment of instances of repeated elements (such as characters) in particular has not been empirically investigated. Additionally, while formalizations describing the role of whole image sections within a sequence have been empirically investigated [9, 11, 12], these models do not systematically address the visual structures within an image section.

3 METHODS AND EXPERIMENTAL SET-UP

3.1 Annotation Scheme

We demarcate several annotation tasks as an attempt to segment comics pages into areas with a meaningful and agreed-upon classification. We focus on identifying rough or sufficient boundaries of areas with significant visual semantic content by having annotators perform the segmentation task itself. We select the most salient elements of typical North American comics for annotation: page segmentation (panels), location reference, character segmentation and reference, and text section segmentation and classification. Each are discussed below, and the annotation scheme in full can be accessed at the link in Appendix A.

3.1.1 Page Segmentation. The *page segmentation* task prompts annotators to delineate the page layout, or the "coherent and distinct image sections" on each page by outlining each segment with a rectangular *bounding box*. Page layout is often synonymous with the arrangement of clearly marked panels separated by empty *gutters*. This scheme, however, does not assume page layout is constrained to panels, but rather emphasises identifying "coherent and distinct" sections.

3.1.2 Location Reference. Annotators are prompted to indicate the depicted or suggested location by assigning a *location reference* label for each page segmentation outlined (e.g. l1, l2...). The location reference remains with a particular location throughout the entire story, meaning that a new label is created only when a new location is perceived. The assigned label therefore acts as a discourse referent, and reflects whether the annotator perceives either repeated or new information regarding the setting per page segmentation.

3.1.3 Text Section Segmentation. Annotators are asked to outline sections of text in the *text section segmentation* task. The use of text to indicate speech utterances, sound effects, or narration is a common feature of comics. In addition, eye-tracking studies suggest that sections of text, in particular those in speech bubbles, receive high rates of fixation [17, 24], indicating their significance to comprehending the narrative.

3.1.4 Text Section Classification. Annotators must *classify the type of text section* for each outlined text section. The options provided are *Speech/thought bubble*, *Narration*, and *Other*, with the latter classification prompting the annotator to input a written description of that section's function.

3.1.5 Character Segmentation. The *character segmentation* task prompts annotators to outline areas depicting agents of narrative focus which play an active role in moving the narrative along. Characters are often depictions of humans (as McCloud notes, humans respond to stories about other humans [23, p. 60]). Other depictions of agents portrayed in a segmentation to set a scene (e.g. crowds) are not considered characters. Similar to text sections, characters are found to have high rates of fixation and are skipped less in eye-tracking studies [17, 24], supporting their role as a fundamental visual element in narrative comprehension.

3.1.6 Character Reference. Annotators are asked to assign a reference label (e.g. x1, x2...) to each outlined character area in the *character reference* task. The assigned character label should be consistent with a character, i.e., the same label should be used for each depiction perceived to be the same character, and hence the first instance of a character is assigned a new label upon introduction. Similar to location reference, the assigned label acts as a discourse referent.

3.2 The Comics Annotation Tool

As the annotation tasks require annotators to demarcate areas on a comic page and assign labels or type classifications to these areas, we created an online tool to accommodate these tasks. The Comics Annotation Tool (CAT) is a browser-based comics mark-up tool, which presents comic pages for annotation as well as responsive elements prompting each annotation task in order. For each segmentation task, annotators outline the approximate size of page segmentations, text sections, and characters directly on the provided comic page image by clicking and dragging a rectangle *bounding box* around the visual element. A new text input form is then

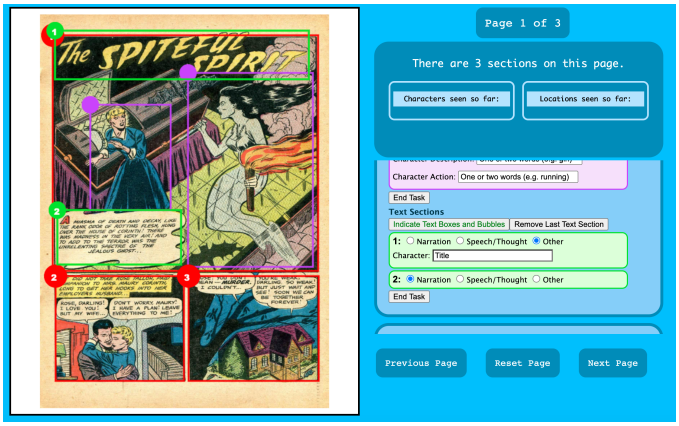

Figure 1: A screen capture of the CAT's main interface.

generated per bounding box in which the annotator may indicate the reference or classification. Figure 1 depicts the main interface - on the left is the comic page being annotated with bounding boxes outlining page segmentations in red, characters in purple, and text sections in green, and on the right are the reference and labelling prompts, matched to their associated by number and color. Only the text section classification task is shown here, with the remaining tasks accessible by scrolling.

Once an annotator completes the annotation tasks for all comic pages in a story, the data is collected in JSON format and sent to an external database.[1]

3.3 Experimental set-up and agreement metrics

3.3.1 Comics Selected for Annotation. Four comic stories with 5 pages each were selected for annotation, making a total of 20 pages annotated. All stories are from the "Alarming Tales" comic magazine, which ran for six issues and was published by Harvey publishers between 1957-8. The comics were downloaded from Comic Book Plus (http://www.https://comicbookplus.com/), which is an internet archive of open source and copy right free comics and similar media. All four stories are in the same genre of fantasy sci-fi. A link to the full comics stories and their publication information can be found in Appendix A.

3.3.2 Participants. A total of ten participants produced the annotations, including the first author. The first author was included as an annotator in order to assess whether the annotation tasks require expert annotators, or annotators highly trained and aware of the theoretical aspects behind the annotation scheme. All other annotators are naive annotators who were only given the annotation scheme and instructions on how to use the CAT to read through before the commencing in annotation.

There were six female and four male participants. All participants were postgraduate students, or friends and partners thereof, recruited from Queen Mary University of London. All participants speak and read English as their native language or to a fluent level.

[1] See Appendix A for a link to access the CAT, full annotation instructions, and code for calculating inter-annotator agreement and generating images on agreement results.

Table 1: Annotator IDs and Total Annotator Pairs per Story

Story Number	Annotators (Numbered)	Total Annotator Pairs
1	0, 1, 2, 3, 6, 7, 8	21
2	0, 1, 2, 4, 8	10
3	0, 1, 2, 5, 8	10
4	0, 1, 2, 5, 9	10

Participants were given Cohn's Visual Language Fluency Index (VFLI) questionnaire to provide a quantitative measure regarding comic reading fluency [10]. All annotators scored in the average fluency range, except annotators 8 and 9 who scored in the low fluency range. The mean VLFI score for all participants is 13.486, indicating average fluency.

Participants were compensated £10/hour worked and could choose the number of stories they wished to annotate, therefore not all stories were annotated by all annotators. Table 1 summarises the annotators per story (the lead researcher is annotator 0) and the total annotator pairs for inter-annotator agreement assessment. All stories have at least 5 annotators, which is a sufficient number to assess for reliability of the annotation scheme.

3.3.3 Inter-annotator Agreement Measures. We use two well-known similarity and agreement measures for image-based and linguistic annotation: *Intersection over Union* for measuring the similarity of bounding boxes outlining elements described in the annotation scheme, and *Cohen's* κ to assess the level of agreement for categorial judgements about those elements.

Intersection over Union (IOU) is a statistic used for measuring the similarity between finite sample sets [25], and is defined as the size of the intersection divided by the size of the union of respective sets A and B: $IOU(A, B) = |A \cap B|/|A \cup B|$. IOU scores are a common evaluation metric for various computer vision tasks. Object detection for images in particular uses IOU to evaluate the accuracy between bounding boxes outputted by an algorithm compared to hand-annotated ground-truth bounding boxes [25, 26, 28], with the resultant IOU score between the bounding boxes providing a quantitative measure of area overlap. A traditional IOU score threshold for a correct object detection is 0.5 or greater [15], though more recent evaluations provide an average over a set of IOU thresholds ranging from 0.5 to 1.00 [18]. A score of 0.7 and above is typically considered a very good score.

The evaluation in this study differs from object detection in that the compared bounding boxes are both hand-annotated, and variation between annotators is expected. However, some segmentation tasks may exhibit greater variation between annotators while still showing adequate amount of overlap to indicate sufficient agreement for rough boundaries of areas with significant visual semantic content. Therefore, a useful outcome of this study is determining satisfactory IOU score thresholds to indicate sufficient agreement according to segmentation task, and provide an overall assessment of the success of a bounding box based annotation scheme.

Cohen's κ is a metric that provides a quantitative degree of agreement between annotators for a classification task, and is widely used in corpus linguistics and other disciplines [2]. The score ranges from

Algorithm 1 Best_IOU_mapping

Input: $ann1StoryData$, $ann2StoryData$

$pages \leftarrow \{\}$ {Initialize dictionary with pages as keys.}
for $page$ from 1 to Length($ann1StoryData$) **do**
  $Ann1 \leftarrow ann1StoryData[page]$
  $Ann2 \leftarrow ann2StoryData[page]$
  $Ann1, Ann2 \leftarrow EqualizeLengthsWithDummys(Ann1, Ann2)$
  $m \leftarrow$ Length($Ann1$)
  $n \leftarrow$ Length($Ann2$)
  Initialise IOU matrix $M[m, n]$ {Populate matrix with IOU scores.}
  for j from 1 to n **do**
    for i from 1 to m **do**
      $M[i, j] \leftarrow IOU(Ann1[i], Ann2[j])$
    end for
  end for
  $BestIOUMean \leftarrow 0$
  $BestMapping \leftarrow \oslash$
  $Permutations \leftarrow P(m, n)$ {All permutations of ann2's annotation indices.}
  for $perm$ in Permutations **do**
    $IOUS \leftarrow <>$
    $Mapping \leftarrow \{\}$
    for i from 1 to m **do**
      $IOUS \leftarrow IOUS + M[i][perm[i]]$ {Retrive IOU score from M.}
      $Mapping[Ann1[i]] \leftarrow Ann2[perm[i]]$ {Add mapping from ann1 to ann2's annotation.}
    end for
    $MeanIOU = Mean(IOUS)$ {Store if better than best permutation so far.}
    if $MeanIOU > BestIOUMean$ **then**
      $BestIOUMean \leftarrow MeanIOU$
      $BestMapping \leftarrow Mapping$
    else
      continue
    end if
  end for
  $pages[page] \leftarrow \langle BestIOUMean, BestMapping \rangle$
end for

Return $pages$

-1 (complete disagreement/negative agreement) to 1 (perfect agreement), with 0 being chance agreement. A score of 0.6 is typically considered adequate agreement, with 0.4 indicating fair agreement and 0.8 and above high agreement. A benefit of Cohen's κ is that it takes into account the chance probability of agreeing based on the frequency of categories in the two sets of annotations. See [7] for details of the calculation.

3.3.4 Mapping Algorithms. While the IOU and Cohen's κ measures described above are suitable for comparing individual pairs of annotations, there may be a different number of annotations made on the same page by two annotators. Therefore, an algorithm is used to compute the IOU measures between all possible pairings between two different sets of annotations on a single page of a certain segmentation task, which then computes the mapping between those two sets of annotations and produces the highest mean IOU score. Pseudocode for *Best_IOU_mapping* is shown in Algorithm 1.

Algorithm 1 takes as input two sets of annotations, one from each annotator being compared. First, if the sizes of these two sets are different, their sizes are made equal by adding a number of dummy annotations, all with the bounding box coordinates $[0, 0, 0, 0]$. This ensures that they will get IOU scores of 0 against any actual annotations, and serves as a punishment within the overall IOU scores for a difference in number of annotations. Second, a matrix is created per page in order to store all the IOU measures between every annotation by annotator 1 against every annotation by annotator 2 on that particular page. Next, all permutations of annotator 2's annotations are tested against a fixed list of annotator 1's annotations. Each of these permutations receives a score, which is the mean of the IOU measures of each element of each of the two lists compared against each other according to their index. Finally, after all the permutations are tried, the one with the highest score and its mapping from annotator 1's annotations to annotator 2's annotations is stored for each page in a dictionary, which is returned when all pages are processed.

Algorithm 2 Best_Cohen_Kappa_reference_score

Input: $IOUMappingStory$, $Ann1Categories$, $Ann2Categories$

$Permuations \leftarrow P(Ann1Categories, Length(Ann2Categories))$ {Get all permutations of ann1's categories.}
$BestCohenKappa \leftarrow -1$
for P in $Permutations$ **do**
  $agreementItems \leftarrow <>$ {Initialize empty list.}
  $RewriteMap \leftarrow \{Ann2Categories_i \leftarrow P_i\}$
  for $page$ from 1 to Length($IOUMappingStory$) **do**
    for $pair$ in $IOUMappingStory[page]$ **do**
      $Cat1 \leftarrow pair.first$
      $Cat2 \leftarrow RewriteMap(pair.second)$
      $agreementItems \leftarrow agreementItems + < Cat1, Cat2 >$
    end for
  end for
  $k \leftarrow CohenKappa(agreementItems)$
  if $k > BestCohenKappa$ **then**
    $BestCohenKappa \leftarrow k$
  else
    continue
  end if
end for

Return $BestCohenKappa$

To calculate the degree to which reference judgements on characters and location are agreed upon, insights from evaluating coreference resolution systems in computational linguistics were used (e.g. [27]), where the emphasis is on correctly identifying co-reference chains (noun-phrases, or discourse referents, in a text which refer to the same the entity), rather than matching labels according to string value alone. Disparate labels for the same intended referent are expected, e.g. one annotator's $x1$ may be consistently labelled what another annotator has labelled $x2$. To avoid classifying this type of string mismatch as a disagreement, we allow for relabelling such that one label assignment identified by one annotator is consistently relabelled with a label assignment from the other annotator's categories. See Algorithm 2 for the pseudocode. The highest score yielded from all possible re-writes is returned.

4 RESULTS

4.1 Segmentation Results

Table 2 reports the mean and standard deviation for the mean IOU agreement scores between each annotator pair for each segmentation task. The overall scores for all annotated segments is shown, in addition to *mapped* segments only, i.e. segments which have mappings to non-dummy segments from Algorithm 1.

4.1.1 Page Segmentation. Overall there is very high agreement between annotators, as the majority of page segmentations were mapped between annotators while also exhibiting high IOU scores.

Table 2: Agreement statistics for all annotator pairs for all segmentation tasks (mapped segments only/mapped and non-mapped segments)

Annotation Task	Story No.	Mean IOU	St. dev.
Page Segmentation	1	0.939/0.931	0.028/0.027
	2	0.955/0.955	0.016/0.016
	3	0.954/0.933	0.010/0.023
	4	0.898/0.784	0.023/0.094
Text Segmentation	1	0.809/0.791	0.036/0.037
	2	0.848/0.836	0.032/0.038
	3	0.764/0.746	0.082/0.086
	4	0.733/0.662	0.071/0.085
Character Segmentation	1	0.725/0.694	0.072/0.091
	2	0.802/0.795	0.041/0.042
	3	0.603/0.538	0.219/0.216
	4	0.657/0.641	0.110/0.104

N=21 for Story 1, N=10 for Stories 2-4

All mean IOU scores were above 0.8 except for Story 4 including non-mapped segmentations, which was just under this at 0.784.

There were several instances of mapping disagreements, mainly within Story 4. Most mapping disagreements occurred on areas of the page exhibiting blocks of text above or next to clearly demarcated panels, where annotators disagreed on whether to consider the text as its own segmentation or to include the text in a segmentation with an adjacent image. Story 4 features more blocks of text outside of demarcated panels than the other stories, resulting in a higher number of segmentation disagreements.

Since the full annotation scheme mentions placing text sections to the right or left of an image in a separate segmentation, and most disagreements regarding text inclusion tended to be typical of a few annotators (primarily 2 and 9), these disagreements appear indicative of whether an annotator followed the scheme closely. Nevertheless, it may be that segmenting out text sections from adjacent images is unintuitive, as the text may directly rely on or supplement the information in the image. Lastly, the IOU scores for such text annexation disagreements are between 0.6 and 0.7, which are still relatively high scores.

There were two cases of disagreements regarding image structure, again both in Story 4. The top areas of the first page of Story 4, which is the top image in Figure 2, was either segmented as one section including the yellow book with text and the title or separated into two sections. This area consistently produced low IOU scores between annotators. For instance, the IOU score between the mapped section between annotators 0 and 9, the former segmenting the yellow book separately with the latter offering one segmentation, is 0.5304. A subsection of the last page in Story 4, which is the bottom image in Figure 2, was either outlined as one segment or split into two separate sections by annotators due to the configuration of the door frame creating a mid-panel demarcation. This area also frequently produced low IOU scores; for instance, the IOU score between Annotators 1 and 2 for the mapped portion of this section is 0.5142.

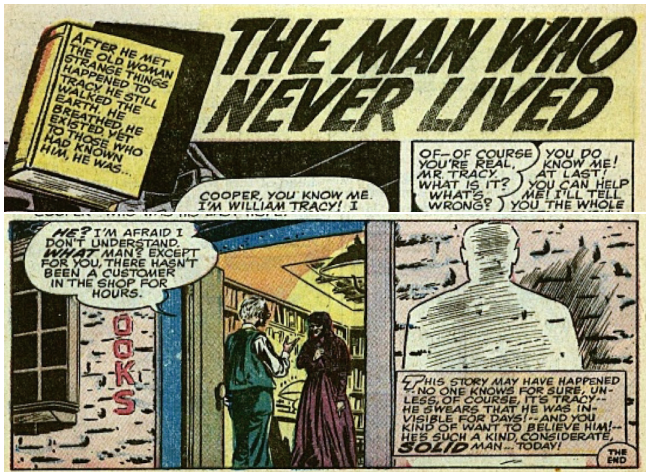

Figure 2: Two subsections from Story 4 exhibiting low page segmentation agreement

4.1.2 Text Section Segmentation. There was high agreement between annotators for text section segmentation, as evidenced by all mean IOU scores being above 0.7 as shown in Table 2, with the exception of Story 4 including both mapped and all annotated text sections having a score of 0.662.

The similar scores on mapped segments only and including non-mapped segments shows the vast majority of text sections were successfully mapped between annotators. Since the mapping algorithm compared text sections between annotators within each mapped page segmentation, Story 4 exhibits lower agreement primarily due to the number of text section annexation disagreements in the page segmentation task, as described in Section 4.1.1. In addition, there were also a few cases of text section disagreement with "the end" sign-offs, which occurred in all stories. Several annotators sectioned out "the end" as separate text sections, while others contained this within larger blocks of text.

Variances in speech bubble outlining were common occurrences across all stories, and could account for most of the low IOU scores between annotator pairs. Speech and thought bubbles tend to be round with tails pointing towards a character to indicate who is speaking or thinking. While annotators were instructed to include the entire speech bubble tail in the bounding box in the full annotation scheme, annotators included the tails to varying degrees across all stories, supporting that the proposed rule appears to be unintuitive to annotators. Marked differences of tail inclusion often resulted in low IOU scores between mapped text sections. Figure 3, for example, depicts an example of relatively low agreement score of 0.5391 for a speech bubble annotation between Annotators 1 and 2, where Annotator 1 cuts off the tail and Annotator 2 includes the whole tail in the bounding box.

4.1.3 Character Segmentation. Table 2 evidences adequate agreement for character segmentation as all mean IOU scores are above 0.6, with the exception of Story 3 including non-mapped segmentations at 0.538. While still showing fairly good agreement, these levels were somewhat below that of page segments and text areas. Figure 4 further visualises the distributions of pair-wise IOU scores

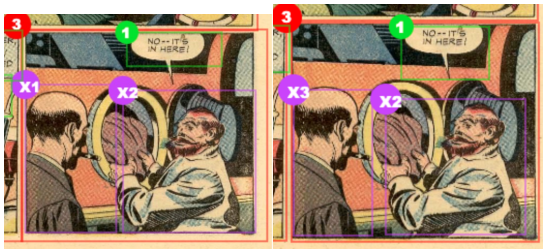

Figure 3: A section from Story 1 with low IOU score for the mapped text section (green bounding box numbered 1) between annotators 1 and 2

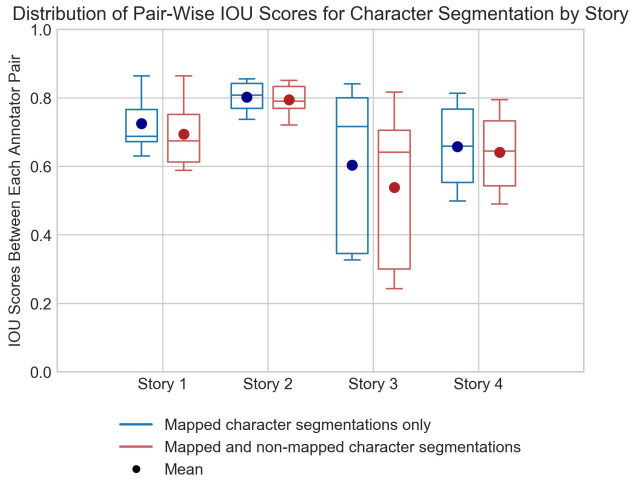

Figure 4: Distribution of pair-wise IOU scores between each annotator for character segmentation per story

per comic story. Each box displays the median and the 1st and 3rd quartiles, while the whiskers depict the maximum and minimum IOU scores per story.

The lower IOU scores appear to be primarily due to the disparate shape of characters, which can be difficult to capture entirely in a bounding box. Additionally, the particularly low agreement for Story 3 can in part be attributed to Annotator 5, who consistently restricted the bounding box to outline just the faces of characters rather than entire character as prompted in the full annotation scheme. Despite lower scores overall, annotators nevertheless do appear to be indicating roughly the same areas on the page.

Most segmentations were successfully mapped between all annotator combinations, meaning that there is high agreement as to which areas on the page depict characters. Consistent disagreement did occur, however, in Stories 1 and 4 regarding agents introduced part-way through the narrative. In Story 1, for example, four of the seven annotators segmented a newly introduced agent but stopped segmenting instances of this agent in later page segmentations, while the other three annotators never segmented the agent. Similarly in Story 4, one of the five annotators did not segment an agent that was depicted in only a few page segmentations, while the four

Table 3: Agreement statistics for all annotator pairs for all reference tasks

Annotation Task	Story No.	Mean Cohen's k	St. dev.
Location Reference	1	0.646	0.187
	2	0.718	0.130
	3	0.919	0.050
	4	0.723	0.109
Text Section classification	1	0.918	0.029
	2	0.959	0.043
	3	0.835	0.138
	4	0.885	0.058
Character Reference	1	0.974	0.019
	2	0.745	0.109
	3	1.0	0.0
	4	0.879	0.104

N=21 for Story 1, N=10 for Stories 2-4

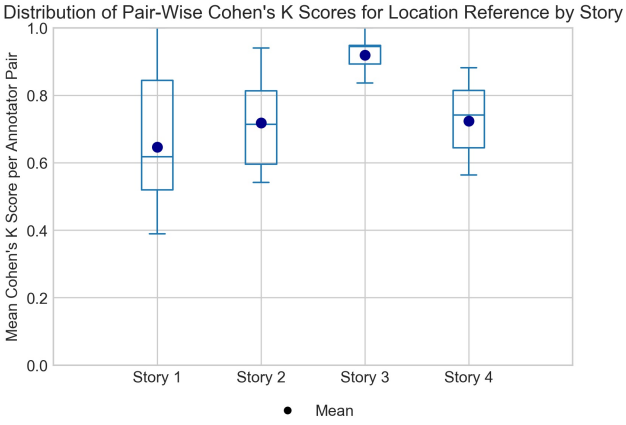

Figure 5: Distribution of pair-wise Cohen's κ scores between each annotator for location reference per story

other annotators consistently segmented this agent. Interestingly, any agent that was shown speaking as indicated by an associated speech bubble was segmented as a character, even if that agent occurred in only one or two image sections.

Lastly, the higher number of segmentation disagreements in Story 3 were due to different interpretations what is considered an active agent. Story 3 featured two men who are attacked by a sentient plant, which is in turn attacked by a robot plant. Annotator 0 segmented the plants as characters, while all four other annotators did not consider the sentient plants to be characters.

4.2 Reference and Classification Results

Table 3 reports the mean and standard deviation for the Cohen's κ scores between each annotator pair for each reference task. Note these results report references for mapped segmentations only.

4.2.1 Location Reference. These results show adequate overall agreement between annotators, as all Cohen's κ scores averaged

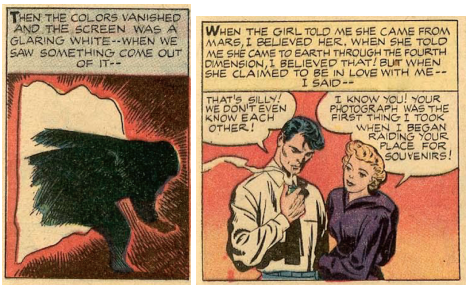

Figure 6: Two (non-sequential) sections from Story 2 exhibiting character coreference with annotation disagreement

above 0.6 as shown in Table 3. However, there appears to be substantial variation in agreement per story, which is further shown in Figure 5 through visualisations of the distributions of pair-wise Cohen's κ scores per comic story.

A small number of disagreements appear to have occurred due to differences in interpretation of location scope. The full annotation scheme specifically states that "there is no need to introduce a new label for small changes in location, such as a character moving to a different corner of the same room - locations should change when *there is a significant change in location or suggested setting*". However, there was disagreement between annotators when characters moved between different rooms in the same building, which seems to occur frequently in Story 1. Three of the seven annotators, for instance, assigned a new reference label to a page segmentation depicting characters looking into another room of the laboratory location, while the other four annotators kept the previous reference label but described the location as a side-room of the laboratory. In addition, disagreements occurred between the first and second page segmentations on the first page of in Stories 1 and 2 - Four of the seven annotators assigned a new referent to the second segment of Story 1, and one of the five annotators did the same for Story 2.

4.2.2 Text Section Classification. The mean and standard deviations of the Cohen's κ score for text type agreement in Table 3 show that there was very high agreement between annotators.

A contributing factor to the lower Cohen κ scores for Story 3 is a number of disagreements between Annotator 5 and all other annotators. Recall that annotators were asked to input a text description of the text section's function. Annotator 5 used the *Other* category to more precisely describe text sections which all other annotators categorized as *Narration*.

Finally, there was some disagreement within the *Other* classification on how to categorize the text that immediately preceded or came after a title. Some annotators categorized this as *Narration*, while others categorized this as "byline" or "tagline" through the *Other* classification. However, annotators agreed on "Title" and "Sound Effect" descriptions in the *Other* category.

4.2.3 Character Reference. The mean and standard deviations of the Cohen's κ score for character reference agreement in Table 3 shows high agreement - even a case of perfect agreement for Story 3. These results show that character reference assignment is high for agreed upon characters segmentations, since these scores only include references for mapped character segmentations.

Similar to location reference, disagreement between the characters occurred between the first and second page segmentations on the first page of Story 1 and 2. Two of the five of annotators for Story 2, for instance, introduced new labels between the character depictions in the first and second page segmentations, while the remaining annotators maintained the previous labels.

Finally, there was also disagreement in Story 2 as to co-reference of a particular character. In this story, a character is introduced that appears as a shadowy figure, shown in the left-side image in Figure 6. Later in the story the character's appearance changes to a woman, and shown in the right-side image in Figure 6. While all annotators initially assigned a new label to the instances of the woman character, three of the five annotators indicated co-reference with the label introduced to the shadowy figure. This co-reference was indicated at different instances of the woman characters, suggesting that annotators recognized this update in character information at different points in the narrative.

5 DISCUSSION

Overall, the results show good to high inter-annotator agreement for the segmentation tasks and sufficient agreement for the reference tasks, evidencing that this approach of hand-segmentation and reference assignment is a promising foundation to build upon. While bounding boxes only give approximate areas of text sections and characters and often include extra space to accommodate surrounding the entire element, they appear to have sufficient agreement for quick implementation without reliance on automatic segmentation programs. Disagreements from reference tasks show that concepts such as "character" and "location", however, do not adequately constrain reader interpretations in all cases.

A useful outcome for future work are IOU score thresholds, which indicate sufficient agreement between annotators per segmentation task for scaling to more complicated comics. From this study, these values are around 0.8+ for page segmentation, 0.6+ for text segmentation and 0.5+ for character segmentation. Additionally, segmentation tasks appear to be a simple task for one or two annotators to mark-up successfully. Reference tasks, on the other hand, appear to require additional verification and cannot be reliably accepted from only one or two annotators.

In terms of disagreements, these divide between those caused by lack of clarify of the current annotation scheme and genuine ambiguity in interpretation. With regards to clarifications to the current annotation scheme, the role of image-based page segmentations that serve to break up the image configuration within a large segmentation requires further investigation. The text section segmentation task description should be amended to instruct the exclusion of speech/thought bubble tails, as this will likely attenuate non-relevant areas in the bounding box. In addition, categories of *title*, *byline*, *sound effect*, and *end salutation* should be explicitly added, and location reference requires a clearer description of scope to minimize variance in agreement across comics stories.

Disagreements evidencing the different reader interpretations suggest re-formulating the ontology to better accommodate ambiguity. First, while page layout is often the fundamental constituent of semantic models of comics [6, 9, 30], the page segmentation disagreements show there is in fact ambiguity in interpreting page

layout, often due to the role of text adjacent to an image. Further work into the *informative content* of text to image, rather than relying on spatial structure alone, is therefore needed. A detailed page layout classification scheme has been developed and tested by [5], however it would be instructive to further consider the content of text and image sections in the context of defining page segmentation, as supported by the results reported here.

Finally, our definition of character produced disagreements in segmentation and reference, revealing that character defined simply as an active agent is not well-defined. There appeared to be particular ambiguity of the role of agents at the beginning of the story, and overall the certainty of agents roles solidified to annotators as the narrative, which is a (perhaps purposeful) ambiguity that draws in the reader [23]. However, there was also disagreement of reference when a character significantly changed form, as well as disagreement on whether a background characters should be segmented consistently or not. From these results we suggest widening the concept of character to that of *agent*, and perhaps prompting annotators to indicate the perceived amount of activity in the story for each given instance of that agent.

6 CONCLUSION AND FUTURE WORK

This study assessed inter-annotator agreement for a preliminary annotation scheme that examines segmentation of areas of comics pages as well as referents and type classifications for these segmentations. It introduced an annotation scheme, a browser-based tool to assist in completing annotation tasks, and methods for inter-annotator agreement assessment. Overall, this approach is a promising foundation to build upon in future work in developing a "bottom-up", empirical ontology of comics.

A limitation of this research is narrow selection of comics. While the artistic style of these comics, which is typical of Silver Age comics, is widespread and exhibits well-known conventions and are therefore likely to present qualities found across a large number of comics, these annotation methods should be tested on a larger number of comics of varying genres and styles. A related limitation is that the comics selected contribute to the very high page segmentation results, as they exhibit mostly rectangular-shaped panels. While the bounding box tool works well for these types of comics, modifications for more precise outlines will most likely be needed for other comics styles.

Finally, there are many ways to build on this work. The relation between information found between text and image for a given section is needed, and a better mechanism to track the evoluability and features of character reference labels can be added and assessed for agreement, as previously discussed. Lastly, additional work must be done to segment or categorize areas within images that are not characters, such as backgrounds and inanimate objects.

ACKNOWLEDGMENTS

We offer a sincere thank you to all the annotators who took part in this study. This research was funded by the EPSRC and AHRC Centre for Doctoral Training in Media and Arts Technology (EP/L01632X/1).

REFERENCES

[1] Dorit Abusch. 2012. Applying discourse semantics and pragmatics to co-reference in picture sequences. (2012).
[2] Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34, 4 (2008), 555–596.
[3] Olivier Augereau, Motoi Iwata, and Koichi Kise. 2017. An overview of comics research in computer science. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 3. IEEE, 54–59.
[4] John A Bateman, Francisco OD Veloso, and Yan Ling Lau. 2019. On the track of visual style: A diachronic study of page composition in comics and its functional motivation. *Visual Communication* (2019), 1470357219839101.
[5] John A Bateman, Francisco OD Veloso, Janina Wildfeuer, Felix HiuLaam Cheung, and Nancy Songdan Guo. 2017. An open multilevel classification scheme for the visual layout of comics and graphic novels: Motivation and design. *Digital Scholarship in the Humanities* 32, 3 (2017), 476–510.
[6] John A Bateman and Janina Wildfeuer. 2014. A multimodal discourse theory of visual narrative. *Journal of Pragmatics* 74 (2014), 180–208.
[7] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.
[8] Neil Cohn. 2013. *The Visual Language of Comics: Introduction to the Structure and Cognition of Sequential Images*. A&C Black.
[9] Neil Cohn. 2013. Visual narrative structure. *Cognitive science* 37, 3 (2013), 413–452.
[10] Neil Cohn. 2014. The Visual Language Fluency Index: A Measure of "Comic Reading Expertise". *Visual Language Lab: Resources* (2014).
[11] Neil Cohn. 2018. In defense of a "grammar" in the visual language of comics. *Journal of Pragmatics* 127 (2018), 1–19.
[12] Neil Cohn and Marta Kutas. 2015. Getting a cue before getting a clue: Event-related potentials to inference in visual narrative comprehension. *Neuropsychologia* 77 (2015), 267–278.
[13] Neil Cohn, Ryan Taylor, and Kaitlin Pederson. 2017. A picture is worth more words over time: Multimodality and narrative structure across eight decades of American superhero comics. *Multimodal Communication* 6, 1 (2017), 19–37.
[14] Alexander Dunst, Jochen Laubrock, and Janina Wildfeuer. 2018. *Empirical comics research: digital, multimodal, and cognitive methods*. Routledge.
[15] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2015. The pascal visual object classes challenge: A retrospective. *International journal of computer vision* 111, 1 (2015), 98–136.
[16] Jochen Laubrock and Alexander Dunst. 2020. Computational approaches to comics analysis. *Topics in cognitive science* 12, 1 (2020), 274–310.
[17] Jochen Laubrock, Sven Hohenstein, and Matthias Kümmerer. 2018. Attention to comics: Cognitive processing during the reading of graphic literature. *Empirical comics research: Digital, multimodal, and cognitive methods* (2018), 239–263.
[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.
[19] Lester C Loschky, John P Hutson, Maverick E Smith, Tim J Smith, and Joseph P Magliano. 2018. Viewing static visual narratives through the lens of the scene perception and event comprehension theory (SPECT). *Empirical comics research: Digital, multimodal, and cognitive methods* (2018), 217–238.
[20] Emar Maier. 2019. Picturing words: the semantics of speech balloons. (2019).
[21] Emar Maier and Sofia Bimpikou. 2019. Shifting perspectives in pictorial narratives. (2019).
[22] Scott McCloud. 1993. *Understanding comics: The invisible art*.
[23] Scott McCloud. 2006. *Making comics: Storytelling secrets of comics, manga and graphic novels*. Harper New York.
[24] Takahide Omori, Taku Ishii, and Keiko Kurata. 2004. Eye catchers in comics: Controlling eye movements in reading pictorial and textual media. In *28th international congress of psychology*. 8–13.
[25] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 658–666.
[26] Adrian Rosebrock. 2016. Intersection over Union (IoU) for object detection. *Online] http://www. pyimagesearch. com/2016/11/07/intersection-overunion-iou-for-objectdetection* (2016).
[27] Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics* 27, 4 (2001), 521–544.
[28] Richard Szeliski. 2020. *Computer vision: algorithms and applications* (2 ed.). Springer Science & Business Media. http://szeliski.org/Book/(visited 2020-12-13).
[29] Mihnea Tufis and Jean-Gabriel Ganascia. 2018. Crowdsourcing Comics Annotations. In *Empirical Comics Research*. Routledge, 85–103.
[30] Janina Wildfeuer. 2019. The Inferential Semantics of Comics: Panels and Their Meanings. *Poetics Today* 40, 2 (2019), 215–234.

A ONLINE RESOURCES

All supplemental materials are available at https://github.com/le300/CAT_Annotation_Experiment_1.