Zeitschrift für
Semiotik
Band 45 • Heft 1–2 (2023)
Seite 57–95
Stauffenburg Verlag Tübingen

Measuring Inter-subjective Agreement on Units and Attributions in Comics with Annotation Experiments

Lauren Edlin and Joshua Reiss, School of Electronic Engineering and Computer Science, Queen Mary University of London

Summary. The conceptualisation of units of interpretation and analyses remains an inherent issue across comics studies. Despite the many conceptualisations of comics units from numerous theories and disciplines, empirical assessments of their validity as proxies for reader interpretation have yet to receive adequate attention. We argue that unit delineation practically involves classifying groups of visual and textual markings according to type, function or semantic category. Based on this, we present a nascent methodology for collecting and measuring inter-subjective agreement by comics readers on proposed units of comics and their attributes. We create an online tool to facilitate handmade segmentations on digital comic pages and assigning labels or classifications appropriate to the annotation task, resulting in segmentation-attribute pairs. We demonstrate the methodology through two inter-annotator agreement experiments that test a segment-attribute pair of panel segmentation and a judgment of background location information. The first experiment shows that assigning a binary classification for panel background judgments requires refinements. The second experiment reconceptualises the task to assess agreement on two scalar methods, namely Likert ratings and a continuous scale. We argue that these experiments support the claim that we can build models of structures in comics with an empirical anchor of reader judgments through this methodology.

Keywords. Comics, visual narrative, corpus annotation, inter-annotator agreement, empirical research methods

Zusammenfassung. Die Konzeptualisierung von Analyseeinheiten bleibt ein grundlegendes Problem der Comicforschung. Obwohl es viele Ansätze zur Differenzierung von Comiceinheiten gibt, wurden die meisten bislang nicht ausreichend empirisch überprüft. In diesem Beitrag argumentieren wir, dass die Abgrenzung von Einheiten aufgrund einer Klassifizierung von Gruppen visueller und textueller Markierungen nach Typ, Funktion oder semantischer Kategorie erfolgen sollte. Hierfür präsentieren wir eine Methodik, die die intersubjektive Übereinstimmung („agreement“) von Comic-Leser:innen bei der Bestimmung von Analyseeinheiten erfasst und prüft. Mithilfe eines Online-Tools ermöglichen wir es, händisch potenzielle Analyseeinheiten auf digitalen Comic-Seiten zu skizzieren und ihnen Attribute und Klassifizierungen in Form einer Annotation zuzuordnen. Wir demonstrieren die Anwendung anhand zweier Experimente, die die Übereinstimmung der Annotationen in der Bestimmung von Segmenten einer Comicseite sowie der Beurteilung von Hintergrundinformationen testen. Das erste Experiment zeigt, dass die Verwendung einer binären Klassifizierung für Hintergrundstandortinformationen unzureichend ist und Verfeinerungen notwendig macht. Das zweite Experiment dient der Bewertung der Übereinstimmung mittels zweier skalarer Methoden – Likert-Skalen und kontinuierlicher Skalen. Diese Experimente, so unsere Argumentation, untersützen die Annahme, dass empirisch verankerte Bestimmungen von Analyseeinheiten durch Comicleser:innen eine Basis für die Erstellung von Modellen für die Struktur von Comics bilden können.

Schlüsselwörter. Comics, visuelle Erzählung, Korpusanmerkung, „Interannotator-Agreement“, empirische Forschungsmethoden

1. Introduction

Comics artists use visual and textual elements to introduce, repeat, emphasise, or de-emphasise information at particular places across a sequence to communicate effectively (Eisner 2008; McCloud 2006). Analysis of comics structures requires precise ways of discussing these elements. A perennial topic in comics studies is therefore conceptualising meaningful and consistent units of interpretation (or analysis), often through developing ontologies to define relationships between units and constrain possible configurations of visual/textual information (Schalley 2019). Current conceptualisations of comics units are typically established within disciplines such as semiotics, linguistics, and computer vision/artificial intelligence. However, empirical assessment of unit conceptualisation validity as proxies for interpretation by everyday readers has not received adequate attention.

Common practices across implicit and explicit approaches towards defining comics units include grappling with the cognitive gap between perceptions of visual markings and higher-level representations, and appropriately delimiting discrete units in non-discrete images and image sequences. In other words, unit delineation involves classifying groups of visual and textual markings according to type, function or semantic category. In this research, we generalise the process of visual element delimitation and classification to develop a practical method for assessing how proposed units are interpreted by multiple readers. Quantitative measures of inter-subjective interpretation can be used in conjunction with current unit conceptualisations as a type of empirical inflection point to further investigate assertions and descriptions about comics structure.

This article presents a nascent methodology to assess inter-subjective interpretation agreement on proposed comics units within an annotation scheme. In general, an annotation scheme prompts annotators to delimit areas on comics pages and assign each area a classification, label, or rating that reflects the type and/or content of the proposed units. The resulting unit is a segmentation-attribute pair. The annotations are assessed using agreement measures commonly used in computer vision and computational linguistics. We demonstrate this methodology's capacity to refine annotation schemes by developing an annotation task, assessing inter-annotator results, and re-conceptualising and retesting the task. This refinement process follows the MAMA (Model-Annotate-Model-Annotate) cycle (Pustejovsky et al. 2017: 24), which is a term coined to express incremental improvement to annotation schemes for text and language corpora. We develop a prototype of a browser-based annotation tool to facilitate efficient annotation of digital comics pages, in which annotators are prompted to create image segmentations paired with classifications regarding the segmentation's content. We investigate the efficacy of this approach through two inter-annotation experiments to test overall reader agreement. We also assess the specific methodological setup, such as investigating expert versus naïve annotators, recruitment through word-of-mouth or crowd-sourcing, and annotation task conceptualisation.

Building directly on our previous work (Edlin and Reiss 2021) which investigated inter-annotator agreement on segmentation and classification tasks assessing panel, character, and text sections, this article offers an in-depth study on interpretations of a specific conception of background location information amount within panel segmentations. Background information is defined as any non-character and non-textual sections of markings in a panel image that explicates the setting or location depicted in that panel, according to a reader. This concept is not only developed to demonstrate the annotation methodology, but is also an attempt to identify places in a comics narrative where the background location appears to have been 'dropped' – that is, the within-panel image does not provide any indication of the location or setting in that panel, and instead typically depicts a single tone. Dropped backgrounds appear across a wide variety of comics types and artistic styles. Figure 1 shows examples of sequential panels from two comics with different art styles and publication formats. Both sequences depict a change in background information amount, namely from some information to no information – in other words, the background has dropped out. The veracity of this concept is tested by reader judgments to determine whether it can be implemented in future work or requires refinement.

The article proceeds as follows: Section 2 gives a brief survey of unit conceptualisations across comics studies to motivate the approach taken in this research. Section 3 describes the Comics Annotation Tool (CAT), which is a prototype browser-based annotation interface used to collect annotations from individual annotators.

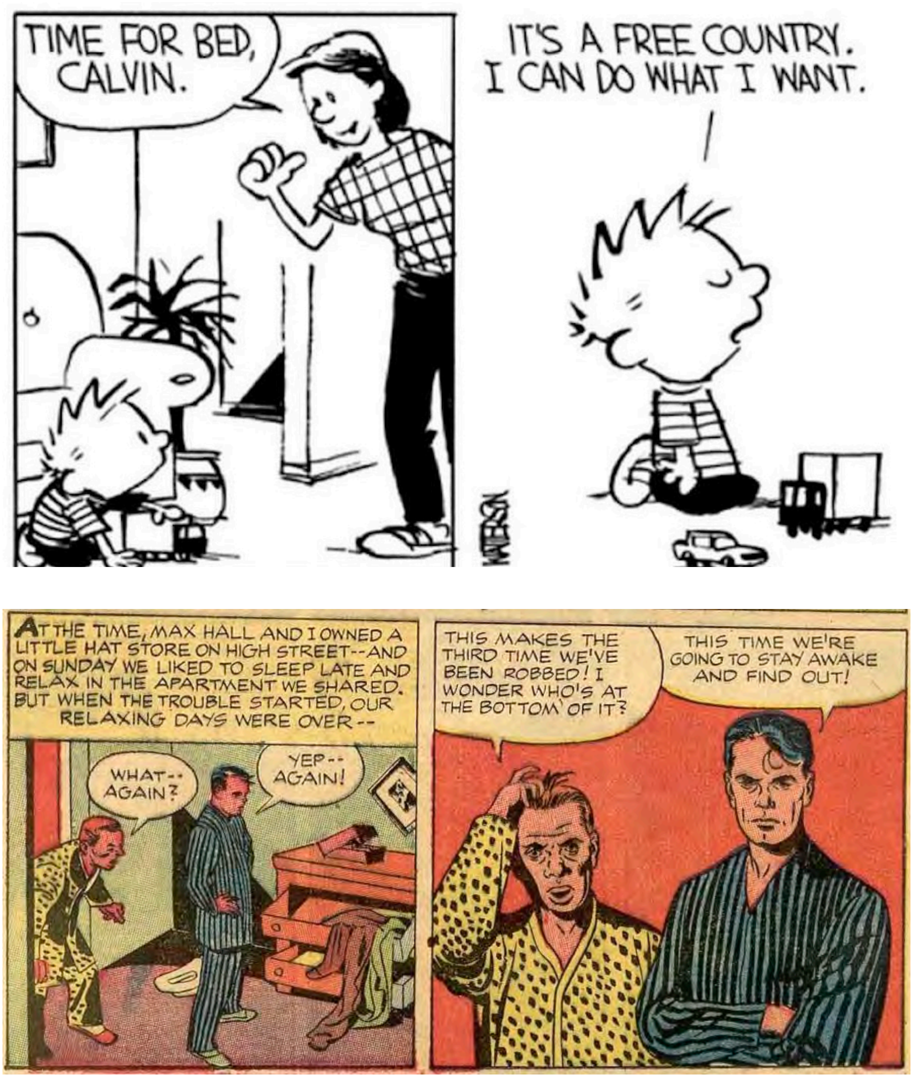

TIME FOR BED, CALVIN.
IT'S A FREE COUNTRY. I CAN DO WHAT I WANT.
AT THE TIME, MAX HALL AND I OWNED A LITTLE HAT STORE ON HIGH STREET--AND ON SUNDAY WE LIKED TO SLEEP LATE AND RELAX IN THE APARTMENT WE SHARED. BUT WHEN THE TROUBLE STARTED, OUR RELAXING DAYS WERE OVER--
WHAT-- AGAIN?
YEP-- AGAIN!
THIS MAKES THE THIRD TIME WE'VE BEEN ROBBED! I WONDER WHO'S AT THE BOTTOM OF IT?
THIS TIME WE'RE GOING TO STAY AWAKE AND FIND OUT!

Fig. 1. Examples of background information amount changes between two sequential panels; **Fig. 1a (top).** Markings in the first panel suggest this scene is located in a living room. No such markings appear in the second panel, although a reader could infer the second panel takes place in the same living room; **Fig. 1b (bottom).** The first panel suggests the action is taking place in a bedroom, however the second panel has a neutral tone besides the character and speech bubbles.

The CAT is used in two experiments described in sections 4 and 5. Section 4 explicates the concepts of panel segmentation and background location information amount, and presents an inter-annotator agreement study where annotators assign a binary classification of some background information, or no background information, to each panel segmentation. Building on these results, section 5 tests several versions of spectrum-based background location amount annotation tasks to refine its previous conception and determine whether readers agree on background information amount using a more fine-grained scale. Reflections on the methodology presented through both experiments are given in section 6. The two experiments demonstrate that the wider background location information conceptualisation needs refinement, although the concept of dropped backgrounds has merit for future work. Through these experiments, the overall methodology is shown to sufficiently capture reader interpretation, although segmentation tasks tend to procure agreement more easily than assignment tasks. Finally, shortcomings of the methodology which point to directions for future work are described.

2. Background: Conceptualisations of units across comics studies

Comics exhibit textual and visual information through repeated conventions and representations. The complexity of these configurations leads to a variety of analytical approaches at various levels of representation, from sub-representational markings (e.g. image contrast, line groups) to high level page compositions (e.g. panel sequences and page layout). An objective comics ontology based on a comprehensive understanding of cognitive processes from visual perception to higher-level semantic representation is not feasible. Many conceptualisations of comics units have therefore been proposed. Such conceptualisations usually follow an established theory or discipline, and are used to investigate or articulate a particular aspect of comics structure and relations between defined elements according to that theory.

We provide a brief, non-exhaustive survey of prominent comic unit conceptualisations from various disciplines, and focus on how units are defined implicitly or explicitly. The methodology described in this research emerges from generalising unit delineation and attribution across these conceptualisations.

2.1 Semiotic and linguistic approaches

Theories and methods from semiotics and linguistics naturally apply to the study of comics, as understanding the relation of form to meaning is foundational to comics research. The backbone of these approaches often relies on defining one or a set of fundamental comics units necessary for the investigation. Therefore, a range of representational levels of units have been defined and analysed.

On the lower end of the spectrum, structuralist scholars examine markings such as lines, dots, or small line groupings and attempt to identify a paired meaning. The inter-relations and compositions of these sub-representational units, in turn, contribute to the meaning of higher-level units such as characters, scenes and panels, and their subsequent paired meanings (e.g. Gauthier 1976: 113). See Cohn (2012), Meesters (2017) and Miller (2017) for summaries of this research. Similarly, the concept of combinatorial morphology formalises how specific combinations of sub-units create coherent compositions. Visual elements that have representational meaning on their own can be affixed with additional markings that produce a novel meaning. For instance, a depiction of a woman with a light bulb above her head indicates that the woman had a sudden burst of inspiration (Cohn 2018b).

More conventional elements such as speech bubbles, instances of characters, and sound effects, are often a primary unit for analysis. Peircean semiotic traditions, which categorise visual signs within complex taxonomies based on a marking's resemblance to its referent, have been widely applied to comics (e.g. Magnussen 2000; Saraceni 2003), and may focus on analyses of a single element. Such investigations include analyses of panels (Caldwell 2012), onomatopoeia (Guynes 2014), characters displaying gestures and action types, and objects (Szawerna 2013), among others. Practitioners of comics also discuss how to effectively implement elements such as speech bubbles and character design in comics creation (Eisner 2008; McCloud 2006). Lastly, units on this representational level are also conceptualised through discourse representation theory (DRT). DRT (Kamp 1981; Kamp et al. 2011) posits that a receiver of a communicative utterance builds a mental representation reflecting information in the communication (Geurts et al. 2020). The mental representation updates accordingly as new information is introduced – however, exactly what information is being updated needs to be defined. Abusch (2012), for example, identifies instances of an individual character as discrete areas in a picture by colouring in the exact shape of the character on page. This work is further built upon by Maier (2019) and Maier and Bimpikou (2019).

A very common primary unit of analysis is panels, which typically grant sequential structure to comics. Panels are also a fundamental unit of the visual narrative grammar (VNG; Cohn 2013a; b; 2018a). Each panel in a sequence is assigned a syntactic category based on its particular narrative function. There are valid and non-valid sequences of panel order according to their narrative category, and valid orders are reflected in abstract hierarchical recursive tree-structures as commonly used in linguistic syntactic analyses (Cohn 2013b: 417; Cohn 2015; Cohn and Kutas 2017). McCloud (1993) describes a taxonomy of six panel transitions based on reader judgments regarding amounts of time and action depicted from one panel to another – the fundamental unit here is a panel dyad, with a judgment on the transition type between them (McCloud 1993). Lastly, panels are a fundamental element with compositions that give rise to page structure, facilitating investigations into relationships between panel structures, panel content, and overall page structure. Groensteen (2007), for instance, argues that panels are the fundamental signifying unit of comics, and coins 'arthrology' as the study of panel configurations, allowing for holistic analyses of sequential and distantly placed panels. Explicit classifications of panel layout compositions lead to further work on the relationship between lower-level panel content and page layout (e.g. Pederson and Cohn 2016; Bateman et al. 2017).

Finally, many analyses examine units and their relations to one another on lower to higher levels of representation. Bateman and Wildfeuer (2014b; a) provide a comprehensive account of higher-level discourse relations within and between panels by applying concepts from Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides 2003). Markings within panels are identified and interpreted through knowledge and experience (e.g. wavy lines above a pipe means a smoking pipe), and demarcated as non-discrete units with a variable assignment. Wildfeuer (2019) further develops delimiting within panel elements by describing a formal notation that assigns perceptually salient features to existential quantifiers and variables. The notation qualitatively expresses entailment between elements that gives meaning to a panel scene, and is a formal description of reader interpretation. Lastly, Yus (2008) describes stages of inferences which a reader goes through when selecting and reading a comic from the cover image to page layouts and processing within and between panel elements.

2.2 Computational approaches

Computational methods are used across comics studies for a wide variety of purposes, which can be broadly subdivided into automatic content identification and automatic content generation. Automatic identification includes detection of sub-representational and conventional elements as well as predicting the presence of an element based on configurations of other elements and features. Automatic generation describes computation-based creation of parts of comics or whole comics. These investigations require precise specifications of the units under analysis (see Augereau et al. 2018; Laubrock and Dunst 2020 for comprehensive surveys).

On the sub-representational level of automatic content identification, computational methods are used to identify and isolate areas of certain textures (Liu et al. 2017) and screen tones (Ito et al. 2015). Properties of images such as colour contrasts, image brightness, the types and numbers of different shapes are detected and used to find higher-level representations such as characters (Mao et al. 2015), or to associate broader concepts such as artistic style (Dunst and Hartel 2018).

Automatic detection of comics features typically involves recognising low-level visual marker configurations, and using techniques from document analysis and computer vision to identify and classify objects in an image. Segmentation tasks seek to match the correct labels to boundary-mark regions of comics images, which are often within a rectangular bounding box containing the sought after visual element. Numerous studies show that text (e.g. Rigaud et al. 2017), speech bubbles and captions (e.g. Dubray and Laubrock 2019), panels (e.g. Pang et al. 2014), and characters (e.g. Nguyen et al. 2017; Qin et al. 2017) can be accurately segmented. Automatic segmentations are compared to a ground truth, or hand-annotated sets of comics pages, to determine a correct segmentation. Knowledge-based ontology approaches add a higher semantic level to assist lower-level extraction processes by adding additional relation constraints between designated elements to facilitate their correct identification. Guérin et al. (2017) describe a formal ontology that uses the concepts of panel, balloon, balloon tail, text line, and character (Guérin et al. 2017: 22) to classify segmentations, or regions of interest, derived from lower-level extraction processes (Rigaud and Burie 2018; Rigaud et al. 2015).

Automatic comics generation, on the other hand, typically describes an ontology of visual and textual elements that a programme selects and organises to create readable comics. The visual and textual elements used to produce comics may be pre-stored – that is, all the drawings and text are already created and not themselves generated – or a series of images is automatically segmented based on some heuristic. The first (to our knowledge) comics generation program, Comic Chat (Kurlander et al. 1996), creates a comic beginning with text from online chat logs. Characters are created by combining pre-drawn heads and bodies, matched with the corresponding text, and placed in a sufficiently sized panel with pre-drawn backgrounds. The placement of character, speech balloons, text elements, and backgrounds on a page are organised according to a set of spatial placement rules. Similar programs inspired by Comic Chat have appeared more recently (e.g. Alves et al. 2008; Soares de Lima et al. 2013; Shamir et al. 2006).

2.3 Proposed approach: Efficient annotation for inter-subjective interpretation measurement and agreement assessment

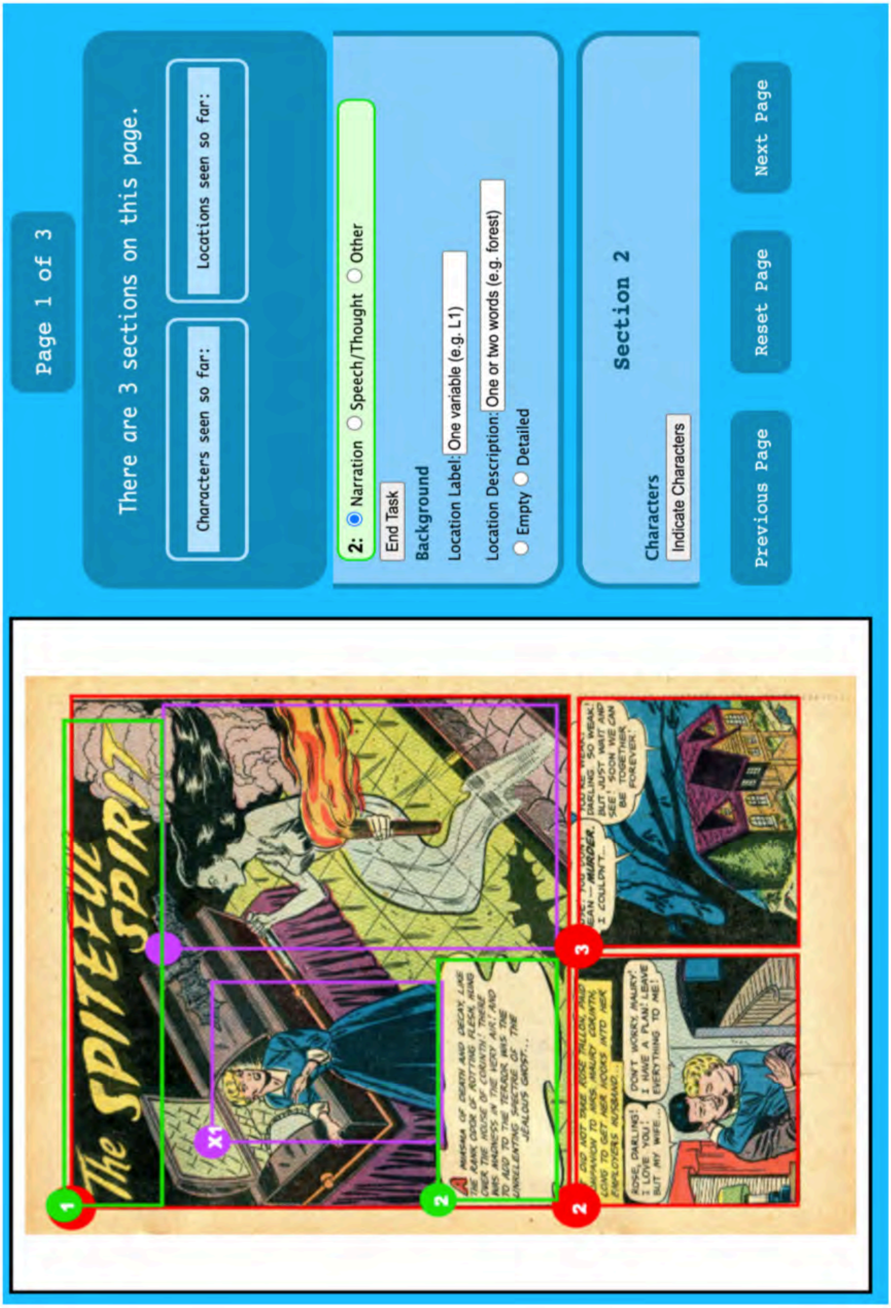
The overview above shows that unit conceptualisation ranges from low to high levels of representation across theoretical and methodological practices. Units are often delineated on different levels within a framework to investigate links from ‘parts to wholes’ in visual content. In other words, each approach grapples with the cognitive gap between perceptions of visual markings and higher-level categories, either by describing how to delimit discrete units from non-discrete images, or by explaining how defined subcomponents contribute to larger visual compositions. The level of unit granularity ranges from high-level image descriptions of panels or page layout as foundations of analysis all the way down to grouping areas of pixels in a comics page image under distinct labels.

What the approaches have in common is that they rely on the judgement of researchers to characterise units, sub-components, and their attributes. What has not received adequate attention is the assessment of inter-subjective agreement by everyday comics readers about the interpretation of the proposed units. Quantifying subjective interpretation is not only useful for understanding whether a unit has been well-conceptualised, but also serves as an empirical grounding for developing links between higher and lower units. Reader interpretation reveals ambiguity and vagueness in particular where readers may have consistently differing interpretations: For instance, if 'character' is a fundamental unit, can we be sure that readers actually discern the same set of markings as representing the same character across a story? Disagreement between readers may reveal intended ambiguity that can be incorporated into a theory, or show that the concept of 'character' is not well-formed.

In light of this, we explore a methodology that facilitates efficient comics annotation from multiple readers to quantify inter-annotator judgments. By abstracting the process of unit conceptualisation from across various methods and approaches, an operational definition of units is achieved by classifying groups of visual and textual markings according to type, function, or semantic category. The annotation process implemented here is therefore developed to have annotators outlining, or segmenting, areas on a comics page and assigning each segment an attribute resulting in a segmentation-attribute pair (segmentations can, of course, have numerous attributes, as tested in one or over several studies). The motivation for the definition of each segment and its attribution is described as a task in an annotation scheme. We borrow several practices and measures from computational linguistics and computer vision to quantity and measure agreement, and follow a MAMA (Model-Annotate-Model-Annotate) cycle (Pustejovsky et al. 2017: 24), where an initial model of units and attributes is created from theoretical assumptions, evaluated, and updated based on the results of inter-annotator experiments. This methodology is demonstrated below through two experiments which test a proposed segmentation-attribute pair unit.

3. The Comics Annotation Tool (CAT)

We aim to assess annotator agreement on the basic amount and type of information across comics pages. The general method we use involves recruiting a number of annotators and providing them with an annotation scheme, which instructs them to demarcate areas of comics pages and assign labels to these segments. We develop the Comics Annotation Tool (CAT) to accommodate these tasks.[1] The CAT is a browser-based comics



Page 1 of 3

There are 3 sections on this page.

Characters seen so far:

Locations seen so far:

2: ● Narration ○ Speech/Thought ○ Other

End Task

Background

Location Label: One variable (e.g. L1)

Location Description: One or two words (e.g. forest)

○ Empty ○ Detailed

Section 2

Characters

Indicate Characters

Previous Page

Reset Page

Next Page

Fig. 2. The CAT's main annotation interface. This version of the CAT is set up for character, text section and panel segmentation on the right, and their associated assignments on the left. Checkboxes for the background information amount task are shown at the bottom of the first light blue section.

mark-up tool that facilitates remote annotation of digital comics pages. Individual annotators can access it via a URL, and are prompted to perform a series of pre-configured annotation tasks in a specific order by instructions and responsive features provided directly in the CAT.

Figure 2 depicts the main CAT interface setup for an experiment testing several segment classification tasks, which are described in previous work (Edlin and Reiss 2021). The comics page on the left is annotated with bounding boxes that are coloured according to the segmentation task.[2] On the right are the reference and labelling prompts, matched to their associated segments on the left by number and colour. Annotators can navigate between pages in the story using buttons at the bottom of the right-hand section of the interface. Once an annotator completes all prompted annotation tasks for all pages in a given story, all segmentations (e.g. pixel positions of bounding boxes on the page) and their associated labels are collected in JSON format and stored in an external database.

When a segmentation task is required, the annotator is prompted to outline areas on the digital comics page image by clicking and dragging rectangular bounding boxes over the desired area. Bounding boxes are used for several reasons: the annotator only has to click and drag on the comic image once, allowing for more efficient and scalable annotation; only two pixel coordinates need to be recorded to assess the size or reproduce the bounding box; and many segmentation tasks for automatic detection of comics elements use bounding boxes, which may facilitate integrating reader-interpreted segmentations described here into corpora that only contain a ground truth, or one interpretation, of the units. Each newly created segmentation generates an input that asks for the annotators' judgment regarding each segmentation's classification, label or reference. The formulation of the input depends on how the labelling task is conceptualised. For example, a binary classification task may provide a checkbox, while a reference labelling task presents a text input.

4. Experiment 1: Inter-Annotator agreement on panel segmentation and binary classification of background information

4.1 Methodology

4.1.1 Annotation scheme and CAT setup

This first experiment investigates panel segmentations and associated background location information, and assesses whether these concepts reach sufficient agreement by aligning readers' interpretations. Panels are conceptualised as coherent and distinct sections based on image structures. Background location information is a judgment about whether visual evidence regarding the location or setting of the narrative is given in a panel.

The agreements we focus on in this paper were carried out in tandem with other annotations tasks (as described and evaluated in Edlin and Reiss 2021). This overall experiment investigated inter-annotator agreement for several proposed annotation tasks regarding paradigmatic comic unit concepts, including panel segmentation and associated location reference, text section segmentation and classification, and character segmentation and reference. A brief overview of the annotation scheme from the overall experiment is provided in Table 1, which specifies the type of label assigned to each type of segmentation, and a summary of the instructions given to annotators.

Tab. 1. An overview of the annotation scheme assessed in the experiment described in Edlin and Reiss (2021).

ANNOTATION	TYPE	SUMMARY OF INSTRUCTION	CAT INPUT
Panel	Segmentation	Outline clear and distinct image sections, even if there is no clear panel outline.	Bounding box
	Reference (Location)	Assign a label that will stay the same for that specific location throughout the story.	Reference Label (e.g. L1, L2)
	Background Information Classification	Indicate whether information about the setting or location is present in a panel background.	Checkbox {Empty, Detailed}
Text Section	Segmentation	Outline any sections of text within a panel.	Bounding Box
	Classification	Indicate the type of text section.	Checkbox {Narration, Speech/Thought, Other}
Character	Segmentation	Outline active participants in the comic narrative within a panel.	Bounding Box
	Reference	Give each character a label that will stay the same for that specific character throughout the story.	Reference Label (e.g. X1, X2)

The concept of background information amount is more precisely defined according to Table 1. Technically, the concept is understood as a judgment regarding the remaining visual markings outside text and character segmentations within a panel segmentation. After creating a panel segmentation with additional text and character segmentations within the panel, annotators were prompted to judge whether there is any information about the wider location or setting in that particular panel without considering markings in the previously segmented text and character areas. If an annotator perceived any such information, they were to select the detailed category. The empty category should be selected if no such information is perceived. Any objects that a character is interacting with are not considered part of a panel background according to the full annotation scheme.

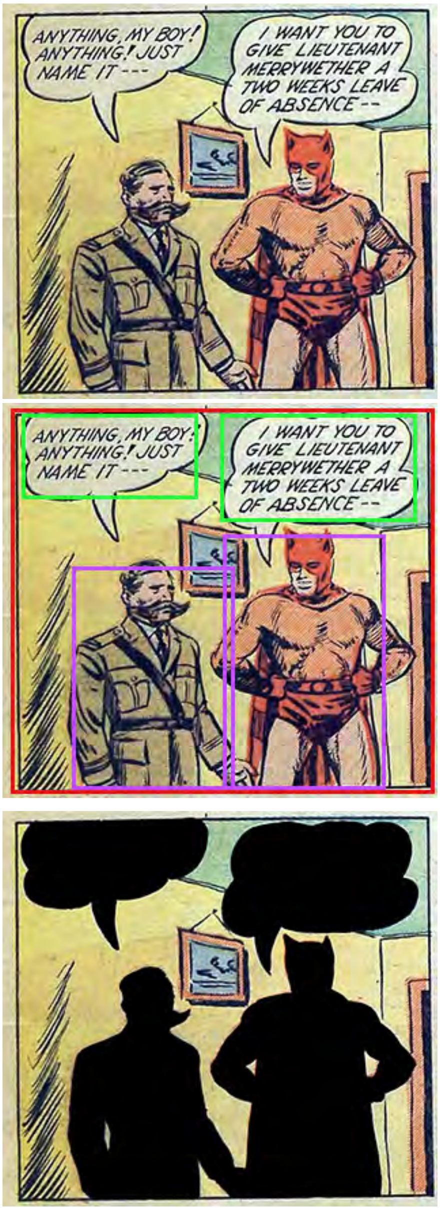

ANYTHING, MY BOY! ANYTHING! JUST NAME IT---
I WANT YOU TO GIVE LIEUTENANT MERRYWETHER A TWO WEEKS LEAVE OF ABSENCE--
ANYTHING, MY BOY! ANYTHING! JUST NAME IT---
I WANT YOU TO GIVE LIEUTENANT MERRYWETHER A TWO WEEKS LEAVE OF ABSENCE--

Fig. 3. A series of panels demonstrating the area meant to be interpreted as the background; **Fig. 3a (top).** The original, unaltered panel; **Fig. 3b (middle).** The same panel with added bounding boxes segmenting text sections and characters; **Fig. 3c (bottom).** The same panel showing only the background areas.

Figure 3 provides an illustrative example of markings intended to be interpreted for background information. Figure 3a depicts a typical panel of a similar style to the comics used in this experiment. Figure 3b shows the same panel with added text section and character segmentations as they would appear in the CAT. Despite the previously described benefits of using bounding box segmentations, they do not precisely outline the complex shapes that often constitute areas depicting text section and characters. Each bounding box includes markings that are technically not meant to be judged for the associated label assignment task. The inverse of this causes some markings that should be included as background segmented within a bounding box. Annotators were therefore instructed to regard bounding boxes as only rough approximations for distinctions between groups of markings intended for interpretation. Figure 3c shows the same panel with a more reasonable expectation of what should be interpreted as background area. The text sections and characters are covered in black, therefore showing only the remaining markings intended for judgment regarding background information amount.

The binary categories of empty and detailed are specifically constructed as an attempt to distinguish panels where all background information is 'dropped' – that is, panels where the location or setting can only be inferred. These types of panels seem to occur across comics from a wide variety of artistic styles, genres, and cultural backgrounds. Refer back to figure 1 in the introduction for two examples. Furthermore, this concept of background location information is constructed to be applicable across many artistic styles; since the annotation task focuses on reader's judgments regarding representations of setting, the actual style of markings – whether an artist uses a richer versus a sparser style, for example – is not a factor in the judgment. However, while the use of dropped backgrounds looks to be prevalent across comics, its use is likely to have different meanings for various authors, genres, or cultural contexts. Identifying cases of these occurrences through reader agreement may therefore be useful in further work to explicate these meanings.

Finally, this unit judgment is used to demonstrate this annotation methodology. As a complex high-level compositional representation, it may be composed of many sub-parts, such as single or groups of objects, other aspects of setting, or more sub-representational aspects such as tones and textures. Since there is incredible potential for image configurations, forcing annotators into a binary choice may reveal instructive disagreements on how to proceed in future work and refine the proposed conceptual unit.

We assessed this annotation scheme on comics stories from *Alarming Tales* comics magazine, published by Harvey for six issues between 1957–1958. We chose these stories because they all have a similar art style that is typical of Silver Age comics, are all of the fantasy sci-fi genre, and are created by several different writers and artists. We limit the scope to one publication to get a precise assessment of agreement on a small set of comparable comics, as disagreements will be difficult to parse on a wider variety of comics at this stage. Finally, these comics have a style that appears to exhibit a range of background information, including potentially clear instances of the 'dropped' background concept. The digital comics were downloaded from the *Comic Book Plus* (2006) Internet archive of comics. Four comics stories with five pages each (for a total of 20 pages) were selected for this experiment.

4.1.2 Participants

A total of ten participants (six female, four male) produced the annotations. All participants are postgraduate students or friends and partners thereof, and were recruited from Queen Mary University of London. All participants speak and read English as their native language or to a fluent level, as all the comics are written in English. Participants were compensated £10/hour and could choose the number of stories they wished to annotate, therefore not all stories were annotated by all annotators. Table 2 lists annotators per story by their ID number and the total annotator pairs for inter-annotator agreement assessment. Annotations made by the first author (annotator 0 in Table 2) are included to assess the effectiveness of naïve versus expert annotators. All other annotators were only given the annotation scheme and instructions on how to use the CAT remotely, and did not receive training or further insight into the experiment.

Readers interpret visual information, including comics images and sequential structures, differently due to cultural background and biases towards particular meanings in visual information. In addition, readers may be more adept at interpretation through more exposure to comics and other visual media. Annotators were therefore given the Visual Language Fluency Index (VLFI) (Cohn 2014) questionnaire which was used to compute a quantitative metric of visual fluency per annotator. All participants scored in the average fluency range, except Annotators 8 and 9 who scored in the low fluency range. The mean VLFI score across all participants is 13.49, indicating average fluency overall.

Tab. 2. Annotator Number and Total Annotator Pairs per Story.

STORY NO.	ANNOTATORS (BY NO.)	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

4.1.3 Inter-Annotator agreement measures

Segmentation agreement was measured using Intersection over Union (IOU, or Jaccard Index). IOU is a quantitative metric of the similarity between two sets, and is defined as the size of the intersection of two sets divided by the size of the union of the same sets: $IOU(A,B) = |A \cap B| / |A \cup B|$. It is a widely used evaluation metric in computer vision (Rezatofighi et al. 2019; Rosebrock 2016; Szeliski 2020). Object recognition in particular evaluates the amount of overlap between an algorithmically generated bounding box against a ground truth bounding box, with the latter surrounding a depicted object attempting to be detected. Figure 4 depicts a visualisation of the IOU metric.

In this experiment, a set is understood as the pixels within a bounding box on a comics page. An IOU score between a pair of annotators' segmentations measures the amount of overlap between both annotator's respective bounding boxes; the more overlap between the segmentations, the higher the IOU score will be. This measurement therefore indicates whether annotators agree on the judgements regarding the use of a comics page's 'space'. For instance, annotators show agreement that the top right of a page shows a panel by segmenting the same areas. Scores are within a range of [0, 1], and what is considered a good score is up to some interpretation. In automatic object detection, a score of 0.5 and above is typically considered a correct detection against a ground truth (Everingham et al. 2015). Previous work (Edlin and Reiss 2021) supports that bounding box segmentations for text agreement should reach a threshold of 0.6, and 0.5 and above for character segmentation.

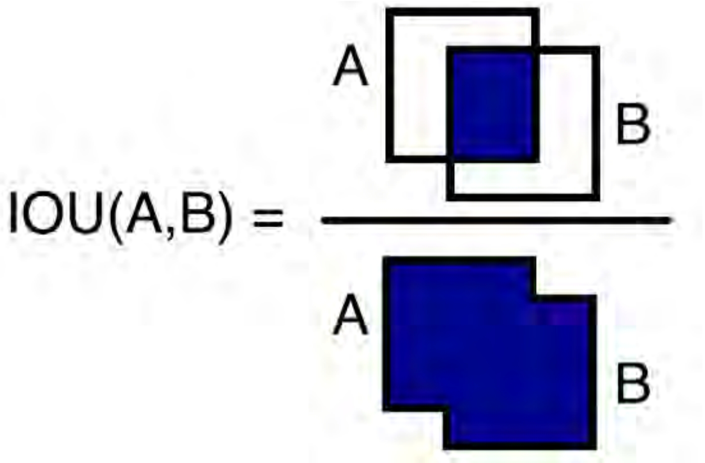


Fig. 4. Visualisation of intersection over union (IOU) between two bounding boxes.

Annotators are instructed to create panel segmentations in the order they read them, and segments are numbered to reflect that order. However, disagreement will typically occur due to a difference in the number of created segmentations between annotators. A mapping algorithm is developed to assess the IOU scores between all possible segmentation pairs between two annotators, per page. The best overall IOU score for a permutation of segmentation pairs gives a 'mapping' of corresponding panels between annotators. The measure of overall panel segmentation agreement is the total mean IOU score of matched panel pairs per annotator, per story. The full mapping algorithm and overall IOU scores that include non-mapped panels are available (Edlin and Reiss 2021). Since the inclusion of non-mapped segments produces a very similar score, only mapped segment results are reported here.

For the background information task, we use Krippendorff's α (KA) (Krippendorff 2011). KA is a quantitative metric of inter-annotator agreement that is widely used in corpus linguistics and content analysis (Artstein and Poesio 2008). The metric rates the extent to which annotators assign the same category or value to the same segmentation. The score ranges from -1 (complete disagreement/negative agreement) to 1 (perfect agreement), with 0 being chance agreement – that is, annotators appear to be randomly assigning values. Although there are no clear thresholds for sufficient agreement, 0.68 is typically considered adequate while 0.8 indicates excellent agreement (Artstein and Poesio 2008: 591). Scores meeting these thresholds indicate a well-conceptualised classification pertaining to content on comics pages, as annotators perform similarly under the same instructions. Lower scores suggest that a task is difficult to understand either due to bad instructions, an incoherent concept, or that annotators are unreliable and choosing random categories.

A KA score is calculated for all annotators against all annotators, as well as between each pair of annotators, per story. Pairwise scores allow for a precise assessment of agreement between each annotator. This gives further context to the all-against-all score, and may indicate potential unreliable annotators. Finally, KA scores were calculated only for mapped segmentations. KA scores were calculated using the *Fast Krippendorf* python package (2017), which is based on the implementation in Grill (2017).

4.2 Results

The results in Table 3 show the means and standard deviations of annotator pair mean IOU scores per story, as well as the mean and standard deviations for annotator pair and all-against-all KAs. The panel segmentations exhibit very high agreement for all stories, except for Story 4, which shows a lower agreement. All scores exceedingly pass the traditional threshold of 0.5, indicating significant overlap between mapped segmentations.

Tab. 3. Panel Segmentation IOU Scores and Background Information KA Scores per Story.

	PANEL SEGMENTATIONS		**BACKGROUND INFORMATION**		
	Pairwise IOUs		Pairwise KAs		All-against-all KAs
	Mean	*Std.*	*Mean*	*Std.*	
Story 1	0.931	0.027	0.3568	0.241	0.3812
Story 2	0.955	0.016	0.526	0.2715	0.5369
Story 3	0.933	0.023	0.6639	0.1347	0.6733
Story 4	0.784	0.094	0.6545	0.1548	0.599

Mapping disagreements show differences in parsing panel structures. There were only a few non-mapped panels in Stories 1–3. These can be attributed to title segmentation – annotators diverged on whether to include large title text within adjacent image segmentations, or place them in their own segmentation. Story 4, however, has many more mapping disagreements, primarily due to different decisions on including text section blocks with neighbouring images. While Stories 1–3 have text section boxes to the left and right of distinct image sections, Story 4 also has text boxes above image sections. The latter composition appears to have a wider range of interpretation in relation to adjacent images. Subsequently, the relationship between areas of text and image in the interpretation of panel should be examined in future work; See our previous work (Edlin and Reiss 2021) for a more detailed discussion.

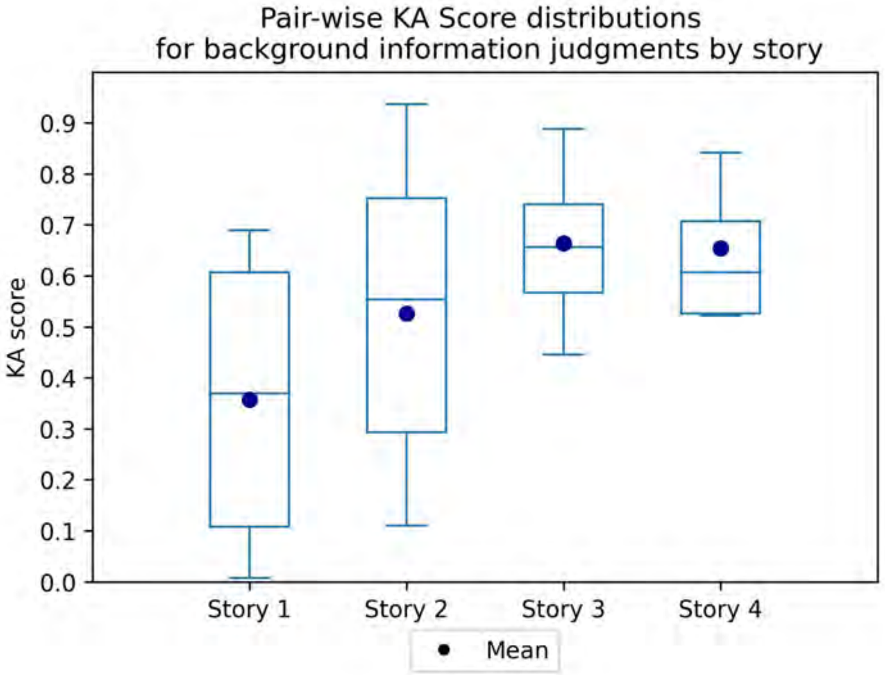


Fig. 5. Boxplots of the distribution of pair-wise KA scores between all annotators for the background location information task, per story.

The background information task generally achieved low agreement across stories, with only the all-against-all KA for Story 3 nearly reaching the 0.68 threshold for adequate agreement. Figure 5 presents the distributions of each individual annotator pair's mean IOU score per story. The pairwise KA distributions show a large range of scores, meaning that some pairs of annotators agreed much more than others. Story 3 exhibits the most overall agreement as it has the highest mean pairwise KA score and the lowest standard deviation between annotators. Story 1 has the least overall agreement, although Story 2 has the highest standard deviation.

4.2.1 Per-panel analyses of background information

Within the 119 mapped panels annotated across all stories, 66 panels (55.5%) exhibit unanimous agreement between annotators, while 53 panels (44.5%) present disagreement from at least one annotator. Story 1 has the highest number of disagreed panels at 74.2%, followed by Story 2 at 54.8%. Stories 3 and 4 exhibit similar percentages of disagreed panels with Story 3 at 22.2%, and Story 4 at 23.3%.

All stories contain unanimously agreed upon empty and detailed panel segmentation assignments. An example of an image that all annotators classified as detailed from Story 4 is shown in figure 6a. These panels often appear to show images from a 'zoomed out' viewpoint. Story 4, followed closely by Story 3, has the most panels with unanimous detailed agreement. Unanimously labelled empty panel segmentations, on the other hand, appear to often depict close-up images of a single character or multiple characters in conversation, accompanied by a solid tone in the remaining space. Figure 6b shows an image from Story 2 classified as empty by all annotators and exemplifies a typical empty panel segmentation. Story 2 has the most panels with unanimous empty agreement and appears to feature many panel segmentations depicting two characters in conversation.

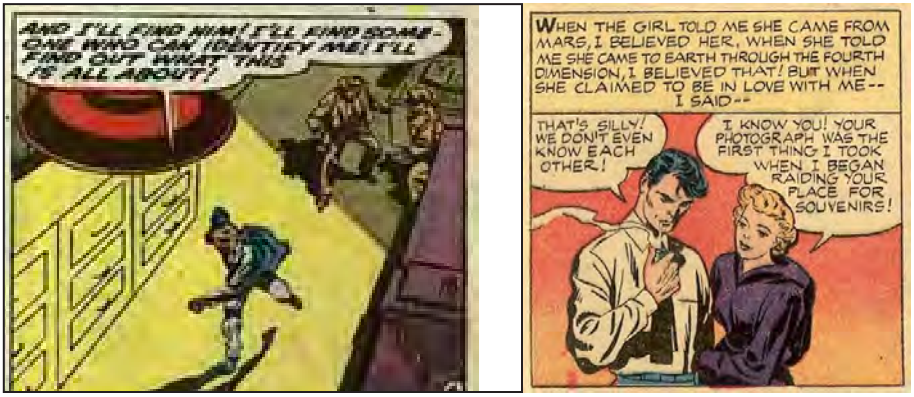

AND I'LL FIND HIM! I'LL FIND SOME-ONE WHO CAN IDENTIFY ME! I'LL FIND OUT WHAT THIS IS ALL ABOUT!
WHEN THE GIRL TOLD ME SHE CAME FROM MARS, I BELIEVED HER. WHEN SHE TOLD ME SHE CAME TO EARTH THROUGH THE FOURTH DIMENSION, I BELIEVED THAT! BUT WHEN SHE CLAIMED TO BE IN LOVE WITH ME-- I SAID--
THAT'S SILLY! WE DON'T EVEN KNOW EACH OTHER!
I KNOW YOU! YOUR PHOTOGRAPH WAS THE FIRST THING I TOOK WHEN I BEGAN RAIDING YOUR PLACE FOR SOUVENIRS!

Fig. 6. Examples of background information classifications with unanimous agreement; **Fig. 6a (left).** A panel segmentation classified as detailed by all annotators (Story 4, Page 4, Panel 8); **Fig. 6b (right).** A panel segmentation classified as empty by all annotators (Story 2, Page 5, Panel 3).

A qualitative assessment of disagreed upon panels reveals several potential causes. Images with discernible and prominent objects in the foreground that also show a single colour or gradient area in the remaining space frequently exhibit disagreement. Figure 7 provides an example. In this image, the scientific instrument is taking up most of the non-character and non-textual space. Since the characters are interacting with the instrument, it should

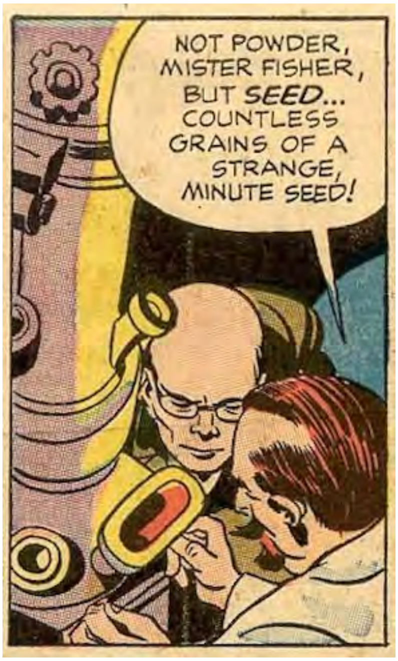

NOT POWDER, MISTER FISHER, BUT SEED... COUNTLESS GRAINS OF A STRANGE, MINUTE SEED!

Fig. 7. An example of image with disagreement possibly due to the image's foreground configuration, which belies the category 'empty' (Story 1, Page 2, Panel 2).

contains a primarily green tone embellished with black shapes. One may infer that these shapes represent a nearby wooded area, meaning that location or setting is present. However, Annotators 1 and 8 classified this section as empty, while Annotators 0, 6, and 7 assigned this section as detailed.

Lastly, disagreements occur when single colour, two-toned, or gradient areas are interpreted to have a meaning relevant to or inferring a wider setting, such as depictions of shadows or sky. Figure 9 depicts two characters looking at two giant intertwined plants, with the remain-technically be discounted from the background. The remainder of the image is a black and blue tone that can be interpreted to be a neutral tone with no location or setting information. This image can therefore be classified as empty according to the annotation scheme. However, most of the image area is taken up by an object, making the category of 'empty' an unintuitive descriptor. Annotators 0 and 7 assigned the segmentation as empty, while Annotators 1, 6, and 8 classified it as detailed. Story 1 seems to exhibit many of these panel types, and is the story with the most disagreed upon panels. Stories 3 and 4 appear to have very few panels with similar prominent foregrounds, and both have the most unanimously agreed detailed judgements.

Second, disagreement commonly occurred for images that appear to show a relatively small indication of the setting, often through only one or two markings. Figure 8 provides an example of such an image. The image background

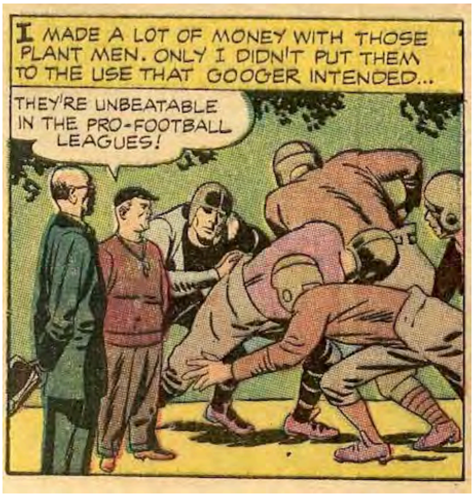

I MADE A LOT OF MONEY WITH THOSE PLANT MEN. ONLY I DIDN'T PUT THEM TO THE USE THAT GOOGER INTENDED...
THEY'RE UNBEATABLE IN THE PRO-FOOTBALL LEAGUES!

Fig. 8. An example of a disagreement based on several background markings (Story 1, Page 4, Panel 6).

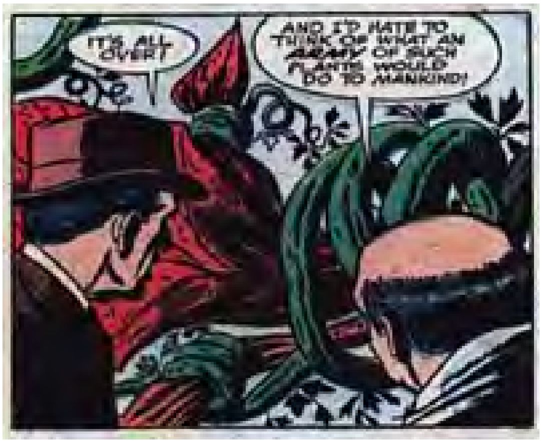

IT'S ALL OVER!
AND I'D HATE TO THINK OF WHAT AN ARMY OF SUCH PLANTS WOULD DO TO MANKIND!

Fig. 9. Example of background disagreement based on interpretations of lighting, sky, or shadows (Story 3, Page 5, Panel 3).

ing non-textual areas filled in with a light blue colour. Annotators 0 and 5 classified this image as detailed, while Annotators 1 and 8 classified it as empty. The blue areas may have been interpreted by Annotators 1 and 8 as a neutral tone to fill the space and therefore assigned empty. Annotators 0 and 5, on the other hand, may have interpreted these areas as the sky, which gives some information on the setting such as being outside in an open area, leading to a classification as detailed. Annotator 1 and 8 may have also interpreted that area as sky, but did not consider the sky to be indicating a more specific or relevant location. In addition, these images tend to have prominent objects presented distinctly in the foreground. The black area in figure 7 above could also be interpreted as a shadow, providing another potential reason for disagreement in this image.

4.2.2 Annotator reliability on background information

Low KA scores are primarily attributed to an incoherent concept for annotation, but can also indicate annotator unreliability. While many annotators tended to agree amongst one another, several annotators consistently disagreed with others across the board. Annotators 3 and 6 in Story 1 consistently produced low agreement, and both were prone to assigning the detailed category against all other annotators empty assignments. Nevertheless, these annotators had an average score on the VLFI metric, which suggests proficient visual language literacy. With these annotators excluded, the all-against-all KA score for Story 1 is 0.5836. It cannot be clearly determined, however, whether these annotators interpreted the instructions in a way conducive to more detailed assignments, or through unreliability. Heatmaps depicting the KA agreement between each annotator pair for each story are available.[3] Finally, there did not appear to be consistent disagreement between expert (Annotator 0) versus all other naïve annotators.

4.3 Discussion

The panel segmentation tasks show very high agreement, while the low to adequate agreement scores for background location information judgments show that forcing a binary choice between two categories is not a robust conceptualisation of background image space. Overall, segmentation tasks proved to yield higher inter-annotator ratings compared to assignment tasks according to previous work (Edlin and Reiss 2021). The panel segmentation results in particular may skew high because most panels in these stories have a rectangular shape that fits well in a bounding box.

A variety of entangled structural and semantic factors contribute to background location information disagreements. Generally, structural disagreements are caused by differences in demarcation of actual background areas, while semantic disagreements occur when the same area is interpreted differently. While some annotators consistently disagreed with all others, disparate interpretations of the task are the main reason for disagreement. The use of the term ‘background’ is likely to have added confusion, as the term may not intuitively align with the image areas intended to be so attributed. Despite these disagreements, the binary categories appear more suitable for some stories than others. It seems that more agreement is accomplished when there is a substantial amount of unanimously agreed detailed panels. This is the case for Stories 3 and 4, both of which had higher scores across the board compared with Stories 1 and 2. None of the stories had a relatively high number of agreed upon empty panels, although it can be speculated that some stories – for instance, comic strips that typically show two characters in conversation – would be suited to a binary classification task.

Nevertheless, the presence of unanimously agreed upon empty classifications across all stories supports that there are image configurations that can intuitively be labelled ‘empty’. This suggests that the panel ‘drop out’ concept is not without merit. We find that these panels tend to have similar visual configurations, and often show one or several characters in conversation with a background tone that does not lend itself to semantic interpretation of setting (e.g. a shadow or sky). It may be beneficial to try and isolate these particular types of images in a refined annotation scheme.

Overall, the low scores for the binary classification and the type of disagreements suggest that the annotation tasks can be improved by taking into account information ‘amount’, or a range of interpretation between no information and some information present. While the survey of potential disagreement sources given above suggests either more structural-based and semantic-based interpretations of background information, these potential causes for disagreement remain intertwined. However, we note that visual configurations that most often produced disagreement have areas of a single colour with additional objects or markings, such as prominent foreground figures or small markings in a space with an otherwise neutral tone. Making additional categories or gradients available to annotators may capture more fine-grained interpretations, without having to specifically address structural and semantic causes of disagreement separately. Gradient scores may also give the description of ‘empty’ a clearer definition. It is unclear, however, what type of scale would most accurately capture interpretations of background information amount.

5. Experiment 2: Crowd-sourced judgements between three scales of background information amount

Building on the results from Experiment 1, we re-conceptualise background location information to be interpreted on a spectrum from no information to full information. We conduct inter-annotator agreement experiments on three different scale types – continuous, ordinal, and binary – to determine which best measures more fine-grained perceptions of background information. We use a between-subjects experiment design and crowd-source a unique set of participants to annotate each scale for each story. Crowd-sourcing comics annotations appears a promising route for gathering large numbers of annotation from a number of annotators (Tufis and Ganascia 2019), therefore we try this approach here.

5.1 Methodology

5.1.1 Annotation scheme and CAT setup

Three versions of the background information amount annotation task are tested: one version prompts annotators to indicate information amount on a continuous spectrum between 1 (no information) and 5 (full information), another version is a 5-point ordinal scale between categories 1 (no information) and 5 (full information), and the final version presents the binary categorical choice of 0 (no information) and 1 (information).[4]

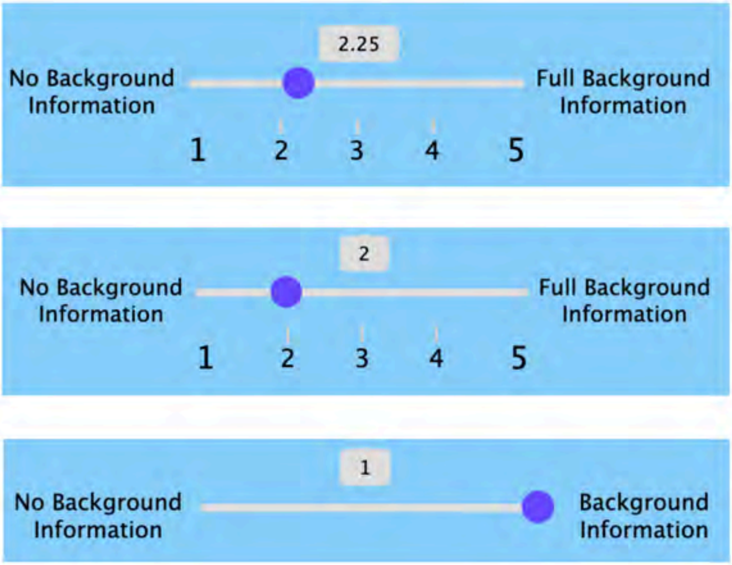

2.25
No Background Information
Full Background Information
1
2
3
4
5
2
No Background Information
Full Background Information
1
2
3
4
5
1
No Background Information
Background Information

Fig. 10. The continuous (top), ordinal (middle) and binary (bottom) scales as presented on the CAT.

Three configurations of the CAT were developed to reflect each annotation scale type. Figure 10 provides an example of how each scale is presented on the right side of the main CAT interface. Annotators can drag the purple circle left and right along the grey line to specify a numerical information amount. The continuous scale on the top allows for indications between integers 1 to 5 up to two decimal places. The ordinal scale in the middle allows the selection of whole integers between 1 and 5. The binary scale allows for the choice between 0 for no information, and 1 for some information. Note that the binary choice task is presented on a scale rather than as a checkbox as in Experiment 1.

Unlike Experiment 1, the comics pages were shown with pre-segmented bounding box panel segmentations to guide annotators to the corresponding section on the right side of the interface. These page segmentations reflect the agreed upon segmentations from Experiment 1, using the lead author's annotations when in doubt.

Stories 1 and 2 from the first experiment are selected for annotation. We use the same stories to allow for a comparison between recruited and crowd-sourced annotations.

5.1.2 Participants

Participants were recruited on the online crowd-sourcing platform *Prolific* (2022). While there are a number of alternative participant recruitment platforms, *Prolific* was chosen for several reasons: i) it was developed specifically for academic research, ii) it could be easily linked to the CAT, iii) it has clear rights and obligations including minimum pay of £5.00/$6.50 per hour, and iv) participants are typically more naïve to experimental research tasks than on other common crowdsourcing platforms (e.g. Mechanical Turk) (Peer et al. 2017).

Tab. 4. Descriptive statistics of participants for Experiment 2.

SCALE TYPE	STORY NO.	NO. COMPLETED (RETURNED /REJECTED)	MEAN AGE	GENDER*	VLFI (MEAN/STD)
Continuous	1	10 (2/0)	26.1	5M/5F	4.9/3.86
	2	10 (6/0)	29.9	1M/8F/1NB	9.8/6.82
Ordinal	1	10 (1/0)	33.1	4M/6F	10.9/6.44
	2	9 (1/1)	30.4	1M/8F	7.75/3.27
Binary	1	10 (7/0)	33.8	3M/6F/1U	5.9/4.17
	2	10 (3/0)	32.2	5M/5F	6.83/5.35

A different set of participants was recruited for each type of scale and each story. Table 4 provides descriptive statistics of these participants, including mean VLFI score per annotator set. Returned participants started the annotation task but did not complete it, and their data was not collected. Rejected participants did not pass the attention check and were also not included in the analysis. Each mean VLFI score per set indicates low fluency, and is below the mean VFLI score for the participants in Experiment 1. Participants were pre-screened through *Prolific* to be between the ages of 21–75, have UK or US nationality, exhibit fluency in English or have English as a first language, and to have achieved an undergrad degree. These attributes were selected to be similar to the demographics of the participants recruited in Experiment 1.

5.1.3 Inter-annotator agreement measures

The new scales produce continuous and ordinal data in addition to the categorical data from the binary scale. KA is used to test for all-against-all annotator agreement as a unifying measure. In addition, we assess the strength of correlation between annotator pairs using Pearson's correlation and Spearman's rank correlation. Spearman's rank is a common method for measuring correlation between Likert scale items, and shows whether the ratings of one annotator correspond to the same rating given by another annotator. Both Pearson's *R* and Spearman's rank correlation coefficients range from -1 to +1, with values close to -1 indicating a negative correlation, values close to +1 indicating a positive correlation, and values near 0 representing no correlation. All calculations were done using the inbuilt functions in the SciPy python library (Virtanen et al. 2020).

5.2 Results

All-against-all KA scores are reported in Table 5. None of the scores reach a sufficient threshold of agreement. However, the binary scale performed better on Story 2 – with results similar to those in Experiment 1 – while the ordinal and continuous scales performed better on Story 1.

Tab. 5. All-against-all KA scores per scale, per story.

	BINARY	ORDINAL	CONTINUOUS
Story 1	0.222	0.5636	0.5033
Story 2	0.5234	0.360	0.3302

Pair-wise agreement was calculated in terms of Pearson's *R* and Spearman's rank for all pairs of annotators for both stories in each condition (binary, ordinal and continuous scales). Overall, the mean of the pairwise correlation coefficients between annotators for each scale type for both stories indicate a moderate mean positive correlation, with the exception of the Story 1 binary scale which exhibits a low positive correlation. Figure 11 displays the distributions of all pairwise correlation coefficients per scale type, per story. For Story 1 in figure 11a, the mean Pearson's correlation coefficients for the continuous (*M* = 0.54, *SD* = 0.16) and ordinal (*M* = 0.63, *SD*

= 0.15) scales, and their respective mean Spearman's coefficients (continuous *M* = 0.61, *SD* = 0.13; ordinal *M* = 0.62, *SD* = 0.16), are all between 0.5 and 0.65. This supports a moderately strong positive correlation. However, the average Pearson's correlation for the binary scale was weak to low-moderate (*M* = 0.27, *SD* = 0.22).

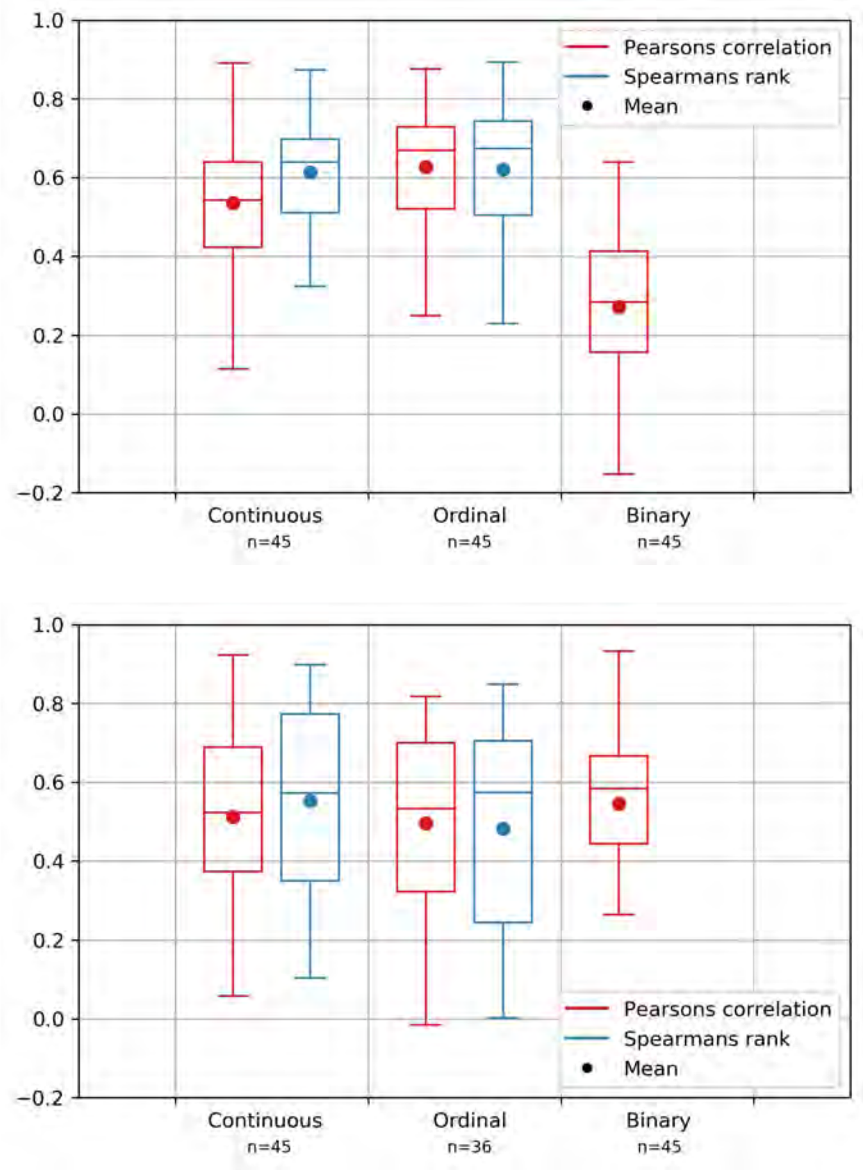


Fig. 11. Distributions of all pairwise correlations coefficients per scale type per story; **Fig. 11a (top).** Story 1; **Fig. 11b (bottom).** Story 2.

For Story 2 in figure 11b, the mean Pearson's correlation coefficients for the continuous (*M* = 0.51, *SD* = 0.22), ordinal (*M* = 0.5 ,*SD* = 0.24), and binary (*M* = 0.55, *SD* = 0.18) scales all indicate a moderate positive correlation. This is also the case for the mean Spearman's correlation coefficients (continuous *M* = 0.55, *SD* = 0.24; ordinal *M* = 0.48, *SD* = 0.26).

In terms of raw means, the continuous and ordinal scales exhibit higher annotator agreement through higher mean pairwise correlation scores than the binary scale for Story 1, while there is little difference between scales for Story 2. To assess whether any of the distributions are statistically significantly different, independent t-tests were performed.[5] An alpha level of 0.01 is used for all statistical tests unless otherwise stated. There are significant differences between the Pearson's correlation score distributions for Story 1 between all three scales: continuous (*M* = 0.54, *SD* = 0.16) and ordinal (*M* = 0.63, *SD* = 0.15) with $t(88) = -2.71$ $p < 0.01$, continuous and binary (*M* = 0.27, *SD* = 0.22) with $t(88) = 6.35$ $p < 0.01$ and ordinal and binary with $t(88) = 8.77$ $p < 0.01$. However, there is no significant difference between the continuous and ordinal Spearman's correlation score distributions for Story 1. There are also no significant differences between the Pearson's or Spearman's distributions for Story 2 between any of the conditions.

5.2.1 Per-panel analyses

To further investigate patterns of higher and lower agreement for panels between and within each condition, we standardise the raw scores per participant by using Z-score transformation and compute the mean and standard deviations of these Z-scores per panel for the continuous and ordinal scales. We compute the percentage agreement for the binary scale.

Within the continuous and ordinal scales, the most disagreed upon panels exhibit similar standard deviations – the five most disagreed upon panels per story for the continuous scale have standard deviations between 0.88–1.1 for Story 1, and between 0.87–1.22 for Story 2. This was similar for the ordinal scale, with Story 1 between 0.83–1.0 and between 0.84–0.96 for Story 2. These panels tend to feature images with objects, scenery and/or characters in the foreground with little visual detail in the background. For the binary scale, Story 1 has three panels that show 50% agreement (the lowest possible score, an even split between annotators) and seven more panels with 60% agreement, while Story 2 only has one panel with 50% agreement and four panels with 60% agreement. The images that exhibit the most disagreement for the binary scale are similar to those that do so for the continuous and ordinal scales.

The most agreed upon panels also produce similar standard deviations for both continuous and ordinal scales. The five most agreed upon panels for the continuous scale have standard deviations between 0.17–0.24 for Story 1, and 0.27–0.44. For the ordinal scale, Story 1 has standard deviations between 0.18–0.24 and 0.17–0.35 for Story 2. The binary scale has five out of thirty-two unanimously agreed upon panels in Story 1, and eleven of thirty-one panels for Story 2. The panels that exhibit high agreement across all scales are overwhelmingly images of characters in conversation against a solid colour background. Exceptions to this occur for several panels; Panel 31 in Story 1 appears to elicit disagreement due to a prominent foreground aspect in the image. Panel 10 in Story 2 shows only a small number of markings in the background which cause disagreements in interpretation. Both these types of disagreements are also found, and further described, in Experiment 1.

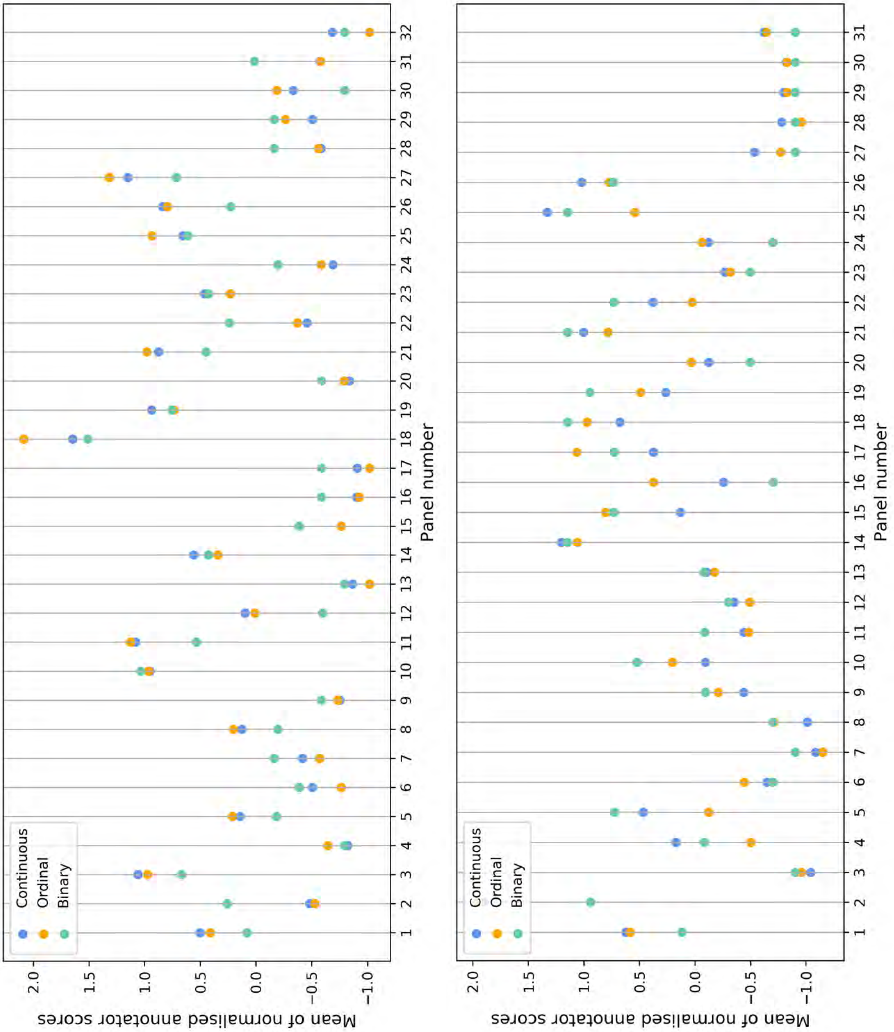


Fig. 12. Mean of annotators values (Z-score normalised) per panel;
Fig. 12a (left). Story 1; **Fig. 12b (right).** Story 2.

The overlap in the most agreed/disagreed panels appears to support generally consistent perceptions of information amounts. We plot the Z-score normalised mean of all annotators scores per scale per panel, which is displayed in figure 12, to further investigate agreement between conditions. The per-panel means for each scale of Story 1 in figure 12a show that the binary scale mean score often deviates from the relatively close ordinal and continuous mean scores, while in Story 2 the binary scale is more closely grouped to the other two scales, as shown in figure 12b.

It also appears that agreement is strongest for panels perceived to be towards the no information end of the scales, and becomes weaker as more background information is perceived. We test this idea statistically by correlating the Z-score normalised mean plotted against the standard deviations per scale, per story, as per figure 13. Pearson's correlations for each scale distribution in figure 13 for Story 1 show that there is a strong positive correlation between mean of annotator score and its standard deviation for the continuous (r(30) = 0.86, p < 0.001), and binary (r(30) = 0.75, p < 0.001) scales, and a moderately strong positive correlation for the ordinal scale (r(30) = 0.55 ,p = 0.001) – this means that the higher the assigned score, the higher the disagreement between annotators – and conversely, the lower the assigned score, more agreement is observed between annotators This relationship is only evident for the continuous scale in Story 2 which shows a low-moderate positive correlation (r(29) = 0.4, p = 0.025, with an alpha level of 0.5), while the ordinal (r(29 = 0.29, p = 0.12)) and binary (r(29) = 0.25, p = 0.18) scales are not correlated significantly.

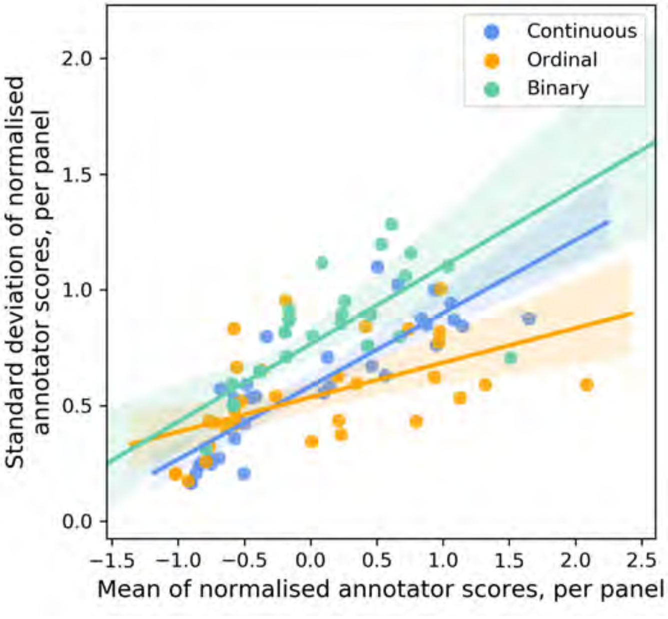


Fig. 13. Z-score normalised mean vs. standard deviation per panel in Story 1.

5.2.2 Annotator reliability

Each condition contained a few annotators who consistently disagreed with most others. However, most annotators were consistent and highly reliable. Therefore, a large range of interpretations of the task by annotators seems to be the primary cause for disagreement. Visual language fluency, as measured by the mean VLFI scores shown in Table 4, does not predict the mean agreement differences in Story 1 – while the participants in the binary condition had a lower VLFI (5.9) than the ordinal condition (10.9), they had a higher VLFI than the continuous condition (4.9), despite having significantly lower mean agreement; for example, see Annotator 3 in the continuous scale condition in Story 2.[6]

5.3 Discussion

Owing to its superior correlations in Story 1, the ordinal scale is tentatively understood as the scale most suitable for achieving better inter-annotator agreement on background information amount when compared to the binary classification. However, the KA results show the background information amount concept does not reach a threshold of agreement for implementation in further comics analysis, even with a finer-grained scale. Similar results for the binary scales between experiments reinforces that reader judgments are consistent between word-of-mouth and crowd-sourced recruitment.

However, the moderate positive mean correlations between annotator pairs for the continuous and ordinal scales convey a consistent relation between annotator perceptions of visual background information amount in a comics panel. The ordinal scale appears to be relatively the most robust, and exhibits the highest inter-annotator agreement for Story 1, but there is no significant difference between the ordinal, continuous and binary scales for Story 2. The binary scale results show a discrepancy in agreement between Story 1 and Story 2, with Story 1 causing more disagreement, consistent with the results from Experiment 1. The KA results for ordinal and continuous scales, however, also show an opposite difference in agreement, with Story 1 exhibiting higher agreement than Story 2. We can conclude that while a particular scale may produce sufficient agreement for stories of a particular style or narrative, no scale is obviously generalisable across comics even within the same publication.

Finally, the notion of the empty or 'dropped' background is again shown to have some conceptual legitimacy, as annotator agreement was highest towards the empty side of the spectrum. This was particularly evident for Story 1, where the ordinal scale showed a less severe increase in disagreement as the mean level of detail increased. This suggests that using a scale of background information amount can be more reliable for empty panel identification but not for gradations of some information – perhaps future work can explore whether the ordinal scale can be used to ‘collapse’ back into a binary classification, where the choice of 1 indicates empty, and a choice of 2–5 indicates non-empty.

6. Discussion

Both experiments, as well as previous work on other segmentation and classification tasks, overall demonstrate that this methodology produces reliable quantified inter-subjective interpretations. The effectiveness of the method is supported in several ways. First, annotators were shown to be more-or-less reliable with only some annotators substantially disagreeing with others. This suggests that disagreements are primarily due to the annotation judgment conceptualisation itself. Second, both experiments produced comparable results for Story 1 and Story 2 using the binary scale. This consistency in annotator judgments across two implementations of the experiment indicates that the practical annotation method itself is sound. It additionally supports the claim that there is no clear benefit between word-of-mouth over crowdsourced recruiting. Panel segmentation with background information amount judgment also seems to be an appropriate task for recruiting everyday readers – a very complex concept requiring theoretical knowledge or expertise is likely to need hand-picked annotators. Segmentation tasks overall tend to attain higher agreement than categorisation or labelling tasks, although agreement thresholds were met for such tasks in previous work. Finally, evidence of minor improvement from the binary to the ordinal scale for the background information location judgement between Experiments 1 and 2 shows that task refinement per the MAMA (Model-Annotate-Model-Annotate) cycle can indeed develop a more robust annotation scheme.

On reflection, the conception of the proposed unit itself determines whether a segmentation-attribute pair is a feasible construction. Sub-representational and lower-level perceptual elements do not appear to be appropriate for this method – recall that many of the disagreements from the experiments were due to different interpretations of background areas with non-representational image components such as neutral tones. However, these disagreements point to places where addressing or incorporating sub-representational elements may be beneficial. Since the overall image area for making a judgment is agreed upon through high-level panel segmentation agreement, perhaps adding a measure of sub-representational marking amounts using computer vision, such as image contrast or number of lines, can be added to panel segments as additional features. Exploring these attributes together may supplement or bolster inter-annotator results – for instance, checking whether correlations of more or less background information in Experiment 2 correspond with certain sub-representational structural features. On the other hand, segmentation can be useful for identifying parts of wholes in higher-level representations. Either way, this refinement methodology provides an initial empirical anchor to the validity of the concept.

7. Future work

There are several shortcomings and limitations with this work that motivate further development of this methodology. First, only a small-sized corpus containing comics exhibiting only one artistic style and genre is used. This limits the findings to a narrow range of comics that were all created at the same point in time, stem from the same cultural background, and show similar artistic styles. Although the findings from the experiments are not generalisable, they do reveal an effectiveness of the presented methodology. Useful disagreements were found even within a small corpus of similar comics, shown here as well as in previous work. Future work may expand on the types and amounts of comics annotated, but it is beneficial to find agreements and parse disagreements in similar stories before implementing the units across comics of different styles, genres, and cultural origins.

Second, the current CAT set-up of implementing bounding box delimiters for annotation are likely to be inappropriate for many segmentations. Bounding boxes were used in this initial implementation to facilitate efficient annotation, and for future comparisons between reader-made segmentations and segmentations found in ground truth corpora for computational analysis. However, many comics use much more diverse panel structures. Many areas of page use, such as sound effects, character outlines, and even sub-representational markings, do not appear to fit well within bounding boxes. A good direction for future work is providing more flexible segmentation tools to the prototype CAT, such as allowing prompting annotators to 'colour in' areas of the comics page as described in (Abusch 2012). In fact, a good direction for future work is developing the CAT to be flexible in allowing researchers to define and implement novel annotation tasks using a set of in-built segmentation and classification tools.

Turning to the background information concept itself, further segmentation or clarification of what area within a panel is meant to be judged is beneficial. Having annotators make a segmentation distinction between traditional understandings of foreground and background, for instance, would help to explain the task. Furthermore, using new terms instead of 'empty', 'detailed' and 'background' may reduce confusion. Nevertheless, since the 'no information' end of the scales in Experiment 2 produced relatively higher agreement, the concept of 'dropped' background has merit to pursue in future work. Eventually applying a robust concept of dropped background to further studies, such as determining their frequency across comics, their meaning, and other relations between use of background space and other features would be informative.

There are also many opportunities for future work using verified segment-attribute pairs using this method. Other examples of how this methodology can be applied include: assessing perceptions along a spectrum of iconicity to representation according Peircean semiotics (how much does a particular visual element resemble what it is signifying, according to annotators?), delimiting areas across a story that make up character discourse chains (Tseng et al. 2018) and other studies of discourse cohesion, and quantitatively modelling amounts of information used by readers to generate meaning from entailment that were qualitatively modelled by Wildfeuer (2019), among others.

Finally, this methodology is conducive to creating comics corpora with multiple reader perceptions per delineated unit. Further analyses of annotator disagreement may indicate cases of deliberate ambiguity or vagueness from the comics' author, and offer a quantified measure of ambiguity. In a practical sense, this methodology produces JSON formatted information about a unit's size and associated attributes according to a number of annotators. This data is therefore useful for further computational analyses of the distribution and constraints of information across comics. A direction for future work is building models which predict selected elements from others. Corpora with empirically verified ground truths could also be useful for automatic content extraction, especially in cases with ambiguity in the comics narrative.

8. Conclusion

The purpose of these two experiments was to provide an example of a general methodology for testing inter-reader interpretations on aspects of comics. The process begins with developing a preliminary concept for a segmentation-classification pair. The concept can be a low, sub-part, or a high-level categorisation, and is typically based on a theory or intuition about comics structure. The concept is translated into an annotation task to test whether everyday readers make the same judgments on areas of comics pages. While the results from previous work showed high agreement for some salient aspects of comics, with segmentation often exhibiting higher agreement than attributions, the background information task discussed here required refinement. The refinement and re-testing follows the spirit of a MAMA (Model-Annotate-Model-Annotate) cycle (Pustejovsky et al. 2017). Achieving high agreement between annotators supports that a proposed unit is well conceptualised, while disagreements are instructive for re-conceptualising stronger units, and may also reveal intended ambiguity or vagueness within the comics narrative. We hope that an efficient annotation methodology allows for developing robust and empirically verified units for future research across comics studies.

Notes

[1] The current version of the CAT is available to use at *GitHub* (https://comicsannotationtool.github.io) and the source code can be downloaded there (https://github.com/ComicsAnnotationTool/comicsannotationtool.github.io) [retrieved April 16, 2024].

[2] The panel segmentations are outlined in red, the character segmentations in purple, and text section segmentations in green. Due to printing constraints the colours are not able to be shown in Figure 2.

[3] See *GitHub* (https://github.com/le300/CAT_Annotation_Experiment_1 [retrieved April 17, 2024]) for heatmaps depicting the KA agreement between each annotator pair for each story.

[4] See *GitHub* (https://github.com/le300/CAT_Annotation_ Experiment_2 [retrieved April 17, 2024]) for the full annotation scheme, which offers detailed instructions and examples.

[5] All assumptions were met to perform independent t-tests between each distribution (data groups are independent, normally distributed, and exhibit similar variance), except for the ordinal scale data for Story 2. This distribution has fewer data points (N = 36) than the others (N = 45) and does not meet the homogeneity of variance assumption using Levene‘s test between ordinal and binary data F(1,79) = 5.74, p = 0.019, and is not normally distributed according to a Shapiro-Wilk test (W = 0.94, p = 0.045). Non-parametric Welch‘s and Mann-Whitney U tests were performed between the distributions for Story 2, with all results showing no significant differences between scale types.

[6] Heatmaps depicting the agreement between each annotator for each story are available at *GitHub* (https://github.com/le300/CAT_Annotation_Experiment_2 [retrieved April 17, 2024]).

Bibliography

Abusch, Dorit (2012). Applying discourse semantics and pragmatics to co-reference in picture sequences. *Proceedings of Sinn Und Bedeutung*, 17, 9–25.

Alves, Tiago, Ana Simoes, Rui Figueiredo, Marco Vala, Ana Paiva and Ruth Aylett (2008). So tell me what happened: Turning agent-based interactive drama into comics. *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems* 3, 1269–1272.

Artstein, Ron and Massimo Poesio (2008). Inter-coder agreement for computational linguistics. In: Hwee Tou Ng (ed.). *Computational Linguistics* 34, 4, 555–596.

Asher, Nicholas and Alex Lascarides (2003). *Logics of Conversation*. Cambridge: Cambridge University Press.

Augereau, Olivier, Motoi Iwata and Koichi Kise (2018). A survey of comics research in computer science. *Journal of Imaging* 4, 7, 87, 1–19

Bateman, John A, Francisco OD Veloso, Janina Wildfeuer, Felix Hiu Laam 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, 476–510.

Bateman, John A. and Janina Wildfeuer (2014a). A multimodal discourse theory of visual narrative. *Journal of Pragmatics* 74, 180–208.

Bateman, John A. and Janina Wildfeuer (2014b). Defining units of analysis for the systematic analysis of comics: A discourse-based approach. *Studies in Comics* 5, 2, 373–403.

Caldwell, Joshua (2012). Comic panel layout: A peircean analysis. *Studies in Comics* 2, 2, 317–338.

Cohn, Neil (2012). Comics, linguistics, and visual language: The past and future of a field. In: Frank Bramlett (ed.). *Linguistics and the Study of Comics*. London: Palgrave Macmillan, 92–118.

Cohn, Neil (2013a). *The Visual Language of Comics: Introduction to the Structure and Cognition of Sequential Images*. London: Bloomsbury Academic.

Cohn, Neil (2013b). Visual narrative structure. *Cognitive Science* 37, 3, 413–452.

Cohn, Neil (2014). The visual language fluency index: A measure of “comic reading expertise”. *Visual Language Lab*. URL: http://www.visuallanguagelab.com/resources [retrieved October 12, 2020].

Cohn, Neil (2015). Narrative conjunction’s junction function: The interface of narrative grammar and semantics in sequential images. *Journal of Pragmatics* 88, 105–132.

Cohn, Neil (2018a). In defense of a “grammar” in the visual language of comics. *Journal of Pragmatics* 127, 1–19.

Cohn, Neil (2018b). Combinatorial morphology in visual languages. In: Geert Booij (ed.). *The Construction of Words*. Cham: Springer, 175–199.

Cohn, Neil and Marta Kutas (2017). What’s your neural function, visual narrative conjunction? Grammar, meaning, and fluency in sequential image processing. *Cognitive Research: Principles and Implications* 2, 1, 1–13.

Comic book plus: No.1 free & legal (2022). URL: http://www.comicbookplus.com [retrieved September 30, 2022].

Dubray, David and Jochen Laubrock (2019). Deep CNN-based speech balloon detection and segmentation for comic books. *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 1237–1243.

Dunst, Alexander and Rita Hartel (2018). The quantitative analysis of comics: Towards a visual stylometry of graphic narrative. In: Alexander Dunst, Jochen Laubrock, Janina Wildfeuer (eds.). *Empirical Comics Research: Digital, Multimodal, and Cognitive Methods*. New York: Routledge, 43–61.

Edlin, Lauren and Joshua Reiss (2021). *An empirically-based spatial segmentation and coreference annotation scheme for comics*. New York, NY: Association for Computing Machinery, 8, 1–8.

Eisner, Will (2008). *Comics and Sequential Art: Principles and Practices from the Legendary Cartoonist*. New York: WW Norton & Company.

Everingham, Mark, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn and Andrew Zisserman (2014). The PASCAL visual object classes challenge: A retrospective. *International Journal of Computer Vision* 111, 1, 98–136.

Fast Krippendorf (2017). URL: https://github.com/pln-fing-udelar/fast-krippendorff.git. San Francisco, CA: GitHub [retrieved November 13, 2022].

Gauthier, Guy (1976). Les peanuts: un graphisme idiomatique. *Communications* 24, 1, 108–139.

Geurts, Bart, David I. Beaver and Emar Maier (2020). Discourse representation theory. In: Edward N. Zalta (ed.). *The Stanford Encyclopedia of Philosophy*. Spring 2020 ed. Stanford: Metaphysics research Lab, Stanford University.

Grill, Thomas (2017). *Python Implementation of Krippendorff's Alpha – Inter-Rater Reliability*. URL: https://github.com/grrrr/krippendorff-alpha.git. San Francisco, CA: GitHub [retrieved November 13, 2022].

Groensteen, Thierry (2007). *The System of Comics*. Mississippi: University Press of Mississippi.

Guérin, Clement, Christophe Rigaud, Karell Bertet and Arnaud Revel (2017). An ontology-based framework for the automated analysis and interpretation of comic books' images. *Information Sciences* 378, 109–130.

Guynes, Sean A. (2014). Four-color sound: A peircean semiotics of comic book onomatopoeia. *Public Journal of Semiotics* 6, 1, 58–72.

Ito, Kota, Yusuke Matsui, Toshihiko Yamasaki and Kiyoharu Aizawa (2015). Separation of manga line drawings and screentones. *Eurographics (Short Papers)*, 73–76.

Kamp, Hans (1981). A theory of truth and semantic representation. In: Paul Portner and Barbara Partee (eds.). *Formal Semantics: The Essential Readings*. Oxford: Blackwell Publishers Ltd, 189–222.

Kamp, Hans, Josef Van Genabith and Uwe Reyle (2011). Discourse representation theory. In: Dov M. Gabbay and Franz Guenthner. *Handbook of Philosophical Logic* 15. Dordrecht: Springer, 125–394.

Krippendorff, Klaus (2011). *Computing Krippendorf's Alpha-Reliability*. URL: http://www.asc.upenn.edu/Krippendorff/2007. University of Pennsylvania: Departmental papers [retrieved October 8, 2022].

Kurlander, David, Tim Skelly and David Salesin (1996). Comic chat. *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*. New York: Association for Computing Machinery, 225–236.

Laubrock, Jochen and Alexander Dunst (2020). Computational approaches to comics analysis. *Topics in Cognitive Science* 12, 1, 274–310.

Liu, Xueting, Chengze Li and Tien-Tsin Wong (2017). Boundary-aware texture region segmentation from manga. *Computational Visual Media* 3, 1, 61–71.

Maier, Emar (2019). Picturing words: The semantics of speech balloons. *Proceedings of the 22nd Amsterdam Colloquium*. Amsterdam: ILLC, 584–592.

Maier, Emar and Sofia Bimpikou (2019). Shifting perspectives in pictorial narratives. In: Uli Sauerland and Stephanie Solt (eds.). *Proceeding of Sinn und Bedeutung* 23, 2, 91–106.

Magnussen, Anne (2000). The semiotics of C. S. Peirce as a theoretical framework for the understanding of comics. *Comics & Culture: Analytical and Theoretical Approaches to Comics* 193–208.

Mao, Xiangyu, Xueting Liu, Tien-Tsin Wong and Xuemiao Xu (2015). Region-based structure line detection for cartoons. *Computational Visual Media* 1, 1, 69–78.

McCloud, Scott (1993). *Understanding Comics: The Invisible Art*. Northampton, MA: Kitchen Sink Press.

McCloud, Scott (2006). *Making Comics: Storytelling Secrets of Comics, Manga and Graphic Novels*. New York: Harper New York.

Meesters, Gert (2017). Semiotics and linguistics. In: Matthew Smith and Randy Duncan (eds.). *The Secret Origins of Comics Studies*. New York and Oxon: Routledge, 100–104.

Miller, Ann (2017). Formalist theory: Academics. *The Secret Origins of Comics Studies*. New York: Routledge, 62–76.

Nguyen, Nhu-Van, Christophe Rigaud and Jean-Christophe Burie (2017). Comic characters detection using deep learning. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* 3, 41–46.

Pang, Xufang, Ying Cao, Rynson W. H. Lau and Antoni B. Chan (2014). A robust panel extraction method for manga. *Proceedings of the 22nd ACM international conference on Multimedia*, 1125–1128.

Pederson, Kaitlin and Neil Cohn (2016). The changing pages of comics: Page layouts across eight decades of American superhero comics. *Studies in Comics* 7, 1, 7–28.

Peer, Eyal, Laura Brandimarte, Sonam Samat and Alessandro Acquisti (2017). Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70, 153–163.

Prolific: A Higher Standard of Online Research (2022). URL: http://www.prolific.co [retrieved September 30, 2022].

Pustejovsky, James, Harry Bunt and Annie Zaenen (2017). Designing annotation schemes: From theory to model. In: Nancy Ide and James Pustejovsky (eds.). *Handbook of Linguistic Annotation*. Dordrecht: Springer, 21–72.

Qin, Xiaoran, Yafeng Zhou, Zheqi He, Yongtao Wang and Zhi Tang (2017). A faster r-cnn based method for comic characters face detection. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* 1, 1074–1080.

Rezatofighi, Hamid, 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. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 658–666.

Rigaud, Christophe and Jean-Christophe Burie (2018). Computer vision applied to comic book images. In: Alexander Dunst, Jochen Laubrock, Janina Wildfeuer (eds.). *Empirical Comics Research: Digital, Multimodal, and Cognitive Methods*. New York: Routledge, 104–124.

Rigaud, Christophe, Jean-Christophe Burie and Jean-Marc Ogier (2017). Segmentation-free speech text recognition for comic books. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* 3, 29–34.

Rigaud, Christophe, Clement Guérin, Dimosthenis Karatzas, Jean-Christophe Burie and Jean-Marc Ogier (2015). Knowledge-driven understanding of images in comic books. *International Journal on Document Analysis and Recognition (IJDAR)* 18, 3, 199–221.

Rosebrock, Adrian (2016). *Intersection over Union (IoU) for Object Detection*. URL: https://pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/ [retrieved December 2, 2020].

Saraceni, Mario (2003). *The Language of Comics*. London: Routledge.

Schalley, Andrea C (2019). Ontologies and ontological methods in linguistics. *Language and Linguistics Compass* 13, 11, 1–19.

Shamir, Ariel, Michael Rubinstein and Tomer Levinboim (2006). Generating comics from 3d interactive computer graphics. *IEEE Computer Graphics and Applications* 26, 3, 53–61.

Soares De Lima, Edirlei, Bruno Feijo, Antonio L. Furtado, Simone Diniz Junqueira Barbosa, Cesar T. Pozzer and Angelo E.M. Ciarlini (2013). Non-branching interactive comics. In: Dennis Reidsma, Haruhiro Katayose and Anton Nijholt (eds.). *International Conference on Advances in Computer Entertainment Technology*. Boekelo: Springer, 230–245.

Szawerna, Michał (2013). A peircean characterization of the semiotic properties of selected visual signs found in comics. *Ars Aeterna* 5, 1, 51–72.

Szeliski, Richard (2020). *Computer Vision: Algorithms and Applications*. 2nd ed. URL: http://szeliski.org/Book/ [retrieved December 13, 2020].

Tseng, Chiao-I, Jochen Laubrock and Jana Pflaeging (2018). Character developments in comics and graphic novels: A systematic analytical scheme. In: Alexander Dunst, Jochen Laubrock and Janina Wildfeuer (eds.). *Empirical Comics Research: Digital, Multimodal, and Cognitive Methods*. New York: Routledge, 154–175.

Tufis, Mihnea and Jean-Gabriel Ganascia (2019). Crowdsourcing comics annotations. In: Alexander Dunst, Jochen Laubrock and Janina Wildfeuer (eds.). *Empirical Comics Research: Digital, Multimodal, and Cognitive Methods*. New York: Routledge, 85–103.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C.J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods* 17, 3, 261–272.

Wildfeuer, Janina (2019). The inferential semantics of comics: Panels and their meanings. *Poetics Today* 40, 2, 215–234.

Yus, Francisco (2008). Inferring from comics: A multi-stage account. *Quaderns de Filologia: Estudis de Comunicacio* 3, 223–249.

Sources

Kida, Fred (art) and unknown (wri) (1957–1958). *My Robot Plants*. New York: Harvey Comics.

Kirby, Jack (wri/art) (1957–1958). *The Cadmus Seed!* New York: Harvey Comics.

Kirby, Jack (wri/art) (1957–1958). *The Fourth Dimension is a Many Splattered Thing!* New York: Harvey Comics.

Oleck, Jack (wri) and Doug Wildey (art) (1957–1958). *The Man Who Never Lived*. New York: Harvey Comics.

Watterson, Bill (wri/art) (1985–1995). *Calvin and Hobbes*. Andrews McMeel Universal syndicate.

Quinlan Sr., Charles (art) and unknown (wri) (1941–1942). *Cat-Man Comics: …Introducing The Kitten*. Helnit Publishing Co. Inc.

Image Sources

Fig. 1a. Watterson (1987). URL: https://www.gocomics.com/calvinandhobbes/1987/06/05 [retrieved May 10, 2024].
Fig. 1b. Kirby (1975: 10).
Fig. 3. Quinlan (1941: 3).
Fig. 6a. Oleck and Wildey (1957, Story 4: 4).
Fig. 6b. Kirby (1957, Story 2: 5).
Fig. 7. Kirby (1957, Story 1: 2).
Fig. 8. Kirby (1957, Story 1: 4).
Fig. 9. Kida (1957, Story 3: 5).

Lauren Edlin, PhD Student
Queen Mary University of London
Electronic Engineering and Computer Science
Peter Landin Building
Mile End Rd, Bethnal Green
London E1 4FZ
United Kingdom
E-Mail: l.edlin@qmul.ac.uk

Prof. Dr. Joshua Reiss
Queen Mary University of London
Electronic Engineering and Computer Science
Peter Landin Building
Mile End Rd, Bethnal Green
London E1 4FZ
United Kingdom
E-Mail: joshua.reiss@qmul.ac.uk