Prepublication version of article to appear in the Handbook of Linguistics and Multimodality

Image Sequences and their Semantics

Dorit Abusch and Mats Rooth

Abstract. This article surveys possible worlds semantics for picture sequences, including comics and film. Topics discussed are geometric semantics for pictures, co-indexing, multimodality, stylization, temporal progression, and embedding.

Keywords: Possible worlds semantics; film; comics; embedding; temporal progression; stylization.

1. Introduction

Artifacts such as pictures, comics, and films are naturally construed as information-bearing, in the same way as sentences of natural language are information-bearing. One approach to theorizing about this adopts a formal model of information content, and analyzes the information content of sentences, comics and film in terms of this single model. The most basic and well-developed account of information content that is applied to natural language is possible worlds semantics, where the information content of a sentence is a set of "possible worlds". A possible world is a world like the one which we inhabit, but where things are more or less different. For instance, at the time of writing in the world we inhabit, there are 193 member states in the United Nations. But we can imagine there being more or fewer, for instance if South Sudan had not yet been admitted. Worlds like that, but which are otherwise similar to our world, are different possible worlds with the same character and specificity as our own. When this assumption is made, the information content of a sentence can be taken to be a set of possible worlds, consisting of the worlds that the sentence is a true description of.[1]

The information content of sentence (1a) is notated as (1b), where the double brackets are Scott/Strachey brackets that map an information-bearing object to its informational or semantic content (Rabern, 2016). (1b) is glossed simply as the set of possible worlds where the number of member states in the United Nations is 193. More technically, the object being mapped to a content is the syntactic structure of a sentence, including lexical items, rather than a mere word sequence.

(1) a. There are 193 member states in the United Nations.
    b. ⟦there are 193 member states in the United Nations⟧

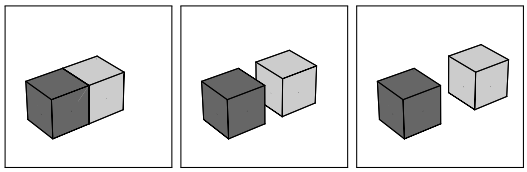
Possible worlds semantics as applied to natural language has been prominent in the philosophy of language and linguistics for more than fifty years (Montague et al., 1970; Lewis, 1972;

This paper is a draft contribution to the *Handbook of Linguistics and Multimodality*, edited by Chiao-I Tseng and John Bateman.

[1]Or depending on the mathematical framework, the information content may be a proper class of possible worlds. A proper class is a collection that includes too much to be a set.

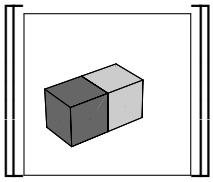
Montague, 1973). Since about 2010, researchers with background in those disciplines have tried to extend it to visual materials such as pictures, comics, and film. This is part of the program of super-semantics, or more generally super-linguistics, where methodology from linguistics and philosophy is applied to studying meaning in non-linguistic materials. Part of the impetus for this comes from multi-modal informational artifacts, such as ones combining pictures with words. To get to a unified semantics for a multimodal artifact, it is helpful and possibly necessary to use the same semantic toolkit for the linguistic and pictorial parts.

To take an example, (2a) is three-panel comic of two cubes moving apart. (2b) is a follow-up in English, which succeeds in combining information from the pictures with information from language. We would like to somehow build up information compositionally, so that the interpretation of the sequence (2a,b) is a set of worlds where a light cube belonging to Amy moves away from a dark cube belonging to Ben. The detailed geometric information in (2a) should put additional constraints on the worlds. A basic step in this is defining a semantics for the left panel in (2a) using the possible worlds toolkit. (2d) writes this possible-worlds "semantic value" for the panel using Scott-Strachey brackets.

(2) a. 

Images Mats Rooth. Abusch and Rooth (2026).

b. The dark cube belongs to Ben and the light cube belongs to Amy.

c. 

This article is organized as follows. Section 2 takes up the problem of defining semantic value like (2c) by a geometric method. Section 3 looks at the analysis of indexing or co-reference in pictorial materials. Section 4 discusses the semantics of multi-modal artifacts. Section 5 looks at issues of time, and Section 6 looks at embedding. Section 7 brings up issues, questions, and alternative approaches, and sums up.

2. Geometric possible worlds semantics

In the early 15th century, Florentine artists developed a geometric method for mapping a three-dimensional scene to a two-dimensional picture (Alberti, 1435; Vasari, 1568). The woodcut in Figure 2, showing two artists making a picture of a lute, describes a physical realization of this mathematical method of perspectival projection (Dürer, 1525). The string is a projection line running from the eyelet at the right to a point on the edge of the soundboard of the lute, where the artist on the left holds the string. The artist on the right positions a stylus at the point where the string intersects the plane of the picture. Then the string is relaxed, the hinged picture is closed, and the artist marks a point on the picture in the position of the stylus. The process is repeated many times to create a line drawing of a lute. This process has been formalized mathematically, and is

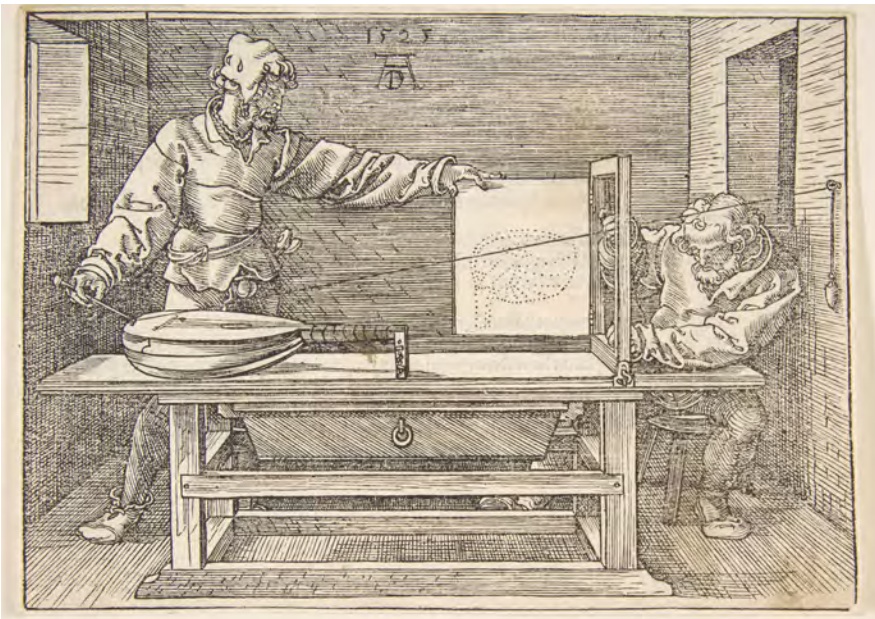


Figure 1: Draftsman Drawing a Lute, by Albrecht Dürer, from his *Instruction in Measurement with Compass and Ruler (Dürer, 1525). This image Dürer (1917).*

realized in software that can create pictures like the comic of the cube in the previous section, or computer-animated films. This process can be conceived of as a function π that maps a world at a time (call it w) and a viewpoint in that world (call it v) to a picture (call it p). Notationally, $\pi(w,v) = p$, which can be glossed as "applying the projection procedure in world w using viewpoint v results in the picture p". Given such a projection function, it can be used to assign semantic values to pictures as sets of possible worlds: the semantic value $[\![p]\!]$ of a picture p is the set of worlds w such that for some viewpoint v, $\pi(w,v) = p$. Less technically, the semantic value of p is the set of worlds that look like p from some viewpoint. This is recorded in (3a). For a number of reasons, it is desirable to keep track of the viewpoint in the semantic value, instead of existentially quantifying it. This is done by taking the semantic value of a picture to be a relation that holds between a world and a viewpoint exactly if the world looks like the picture from the viewpoint. This is recorded in (3b), where the relation is described as a set of ordered pairs.

(3) a. $[\![p]\!] = \{w | \exists v.\pi(w,v) = p\}$
b. $[\![p]\!] = \{\langle w,v\rangle | \pi(w,v) = p\}$

One reason for preferring the second version is that constraints on successive viewpoints are often imposed or inferred. It is natural to construe the comic with the cubes as assuming a constant viewpoint, with the dark cube understood as motionless, and the light cube understood as moving away from it. Section 6 on temporal progression will analyze this comic in terms of three verifying worlds w_1, w_2, w_3, each a temporal extension of the previous one. w_1 looks like the first picture from a viewpoint v_1, w_2 looks like the second picture from a viewpoint v_2, and w_3 looks like the third picture from a viewpoint v_3. If there is access to the viewpoints via semantic values as in (3b), then the information that the viewpoint stays constant can be added when information from

Figure 2: Stills from a five-shot film of a chess game. Viewers infer that the dark-haired man is playing white and the blond man playing black. Since no shot shows both the chess board and a player, the inference must come from constraints on the viewpoints for successive shots. Images from Cumming et al. (2013), Cumming et al. (2017).

the three panels is combined, $v_1 = v_2 = v_3$. A similar point can be made for film. Cumming et al. (2017) studied constraints on viewpoints for successive film shots. They argue that many situation types depicted in film obey an "X-constraint", where the camera position stays on one side of an "action line". This holds for instance for a scene with two characters conversing across a cafe table, where the camera stays on one side of a line connecting the two characters. As an example, Figure 2 shows stills from a short film of chess game, consisting of five shots. Cumming et al. report that viewers infer that the dark-haired man is playing white, and the blond man is playing black. Since none of the shots show both the chess board and a player, the inference must come from the assumption that the X-constraint is obeyed.[2] Stating the X-constraint in terms of semantic values for successive shots requires access to viewpoints in the semantic values, not just worlds.

The idea found in current work that pictures have a propositional semantic value like (3a) comes from Gabriel Greenberg's Ph.D. dissertation (Greenberg, 2011). Of course, the mathematics of perspectival projection traces back for half a millennium. The impulse to re-interpret it as possible worlds semantics for pictures comes from the experience of researchers in the philosophy of language and linguistic semantics, who work with possible worlds semantics as a minimal workable model of information content. Equally, it comes from the experience of readers of novels, viewers of film, and readers of graphic novels who experience these media as conveying narrative information that is strikingly similar in many respects. Finally, as will be seen in Section 4, when working on the semantics and pragmatics of multimodal artifacts such as comics with captions, speech bubbles, and thought bubbles, or film with intertitles, it is desirable to use the same semantic toolkit for linguistic and pictorial materials.

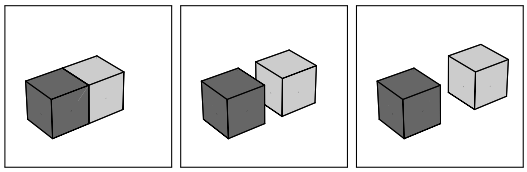
3. Co-indexing

The three panels in the comic repeated in (4) can be given independent semantic values, a set of pairs of worlds and viewpoints. These should be pieced together into a unitary semantics, which places constraints on a single world. One aspect of this is temporal progression, where time advances in the described world from one panel to the next. This is discussed in Section 6. There is another issue of the identity of depicted objects across panels. The natural construal of (4) has it that the dark cube depicted on the left in the first picture is the same object as the dark cube depicted on the left in the second and third pictures. Similarly for the light cubes on the right.

[2]Since the information in a film shot extends in time, the semantic setup has to be more complicated. Rooth and Abusch (2026) say that the information in a film shot is a set of pairs $\langle \vec{w}, \vec{v} \rangle$, where $\vec{v}$ is a sequence of viewpoint/camera positions, and $\vec{w}$ is a sequence of worlds at a time, with each world temporally extending the previous one.

But the sequence of pictures as reflected in their geometric semantics is consistent with a scenario where, in the time between the times that verify the first and second pictures, a dark cube zips out of view, and is replaced by a different dark cube. The reason is that two different cubes of identical shape and color can look the same from a viewpoint. The basic geometric semantics, then, is non-committal whether the cubes depicted in the panels are the same object or not.

(4)

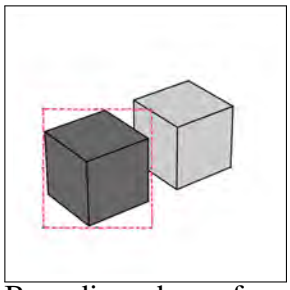


In order to capture the obvious interpretation, a mechanism is needed for identifying depicted objects across pictures. To achieve this, something has to be changed in the semantic values, because an unstructured relation between worlds and viewpoints does not give access to depicted objects. This situation is analogous to the phenomenon of cross-sentential anaphora in language. Suppose the first sentence in (5a) is interpreted in possible worlds semantics as a set of worlds—the ones that satisfy the constraint involving the old oak. While this proposition consists of worlds where there is an oak of age more than four hundred years in New York state, it does not give any direct access to the oak in a given world of the proposition. This makes it difficult to interpret the subject pronoun in the second sentence. It would be better if the semantics of the first sentence gave access to a set of *pairs* $\langle w,x\rangle$ such that in w, x is an oak in New York state of age more than 400 years, or some semantic value that records such a choice for the oak. This is exactly the approach taken in dynamic semantics and discourse representation theory (Lewis, 1975; Kamp, 1981; Heim, 1982). Indefinite descriptions like *an oak* set up discourse referents, and in a pair $\langle w,x\rangle$, x functions as a witness for the discourse referent. In some versions of dynamic semantics, the semantic value of (5a) is exactly a relation between worlds and witnesses for the oak (Dekker, 1994, 2012). This semantic value makes it possible to interpret the pronoun in (5b).

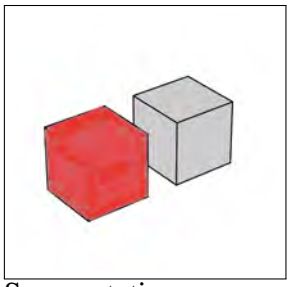
(5) a. In New York state, there is an oak that is more than four hundred years old.
b. It has a diameter of more than two meters at a distance from the ground of two meters.

The above strategy of analysis has been duplicated directly in the analysis of indexing in pictorial narratives, using either discourse representation theory (DRT) or dynamic semantics (Abusch, 2012). When interpreting natural language, one can say that indefinite descriptions, which are a certain kind of syntactic phrase, trigger the introduction of discourse referents. In a basic picture, there is nothing like a certain kind of sub-constituent or feature that can be held responsible for the introduction of discourse referents for depicted objects. Instead something has to be added. A similar problem comes up in machine vision and artificial intelligence, where there is an issue of how to add a predication like $\mathbf{cube}(x)$ or $\mathbf{cat}(x)$ to a picture, where x is a depicted object. Two ways of doing this are bounding boxes and segmentation maps (Lempitsky et al., 2009). (6) illustrates this for one of the cube pictures. A bounding box is a geometric rectangle in the picture plane that includes the projection of the depicted object, here a cube. A segmentation map marks those points in the picture plane (here in red) that are within the projection of the cube. Finally, at the price of increased ambiguity, one can use a single point from the segmentation map to introduce the discourse referent.

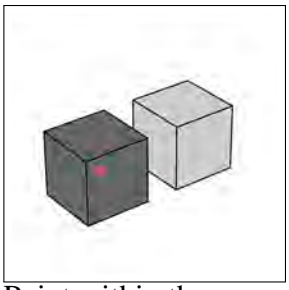
(6)



Bounding box for the projection of the dark cube.



Segmentation map with area marked in red.



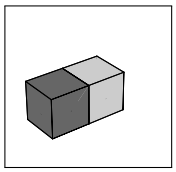
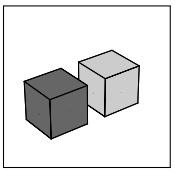
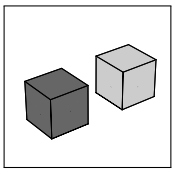
Point within the projection of the dark cube.

Images ©ⓘMats Rooth. Abusch and Rooth (2020-2026)

.

The idea in the three options is the same. A projection procedure not only maps a world w and a viewpoint v to a picture p, it establishes a correspondence between objects in world w that are in view from v and areas of p. Points have the advantage of allowing a simpler characterization of the witnesses for the discourse referents. Notationally, introductions of discourse referents using points are interleaved into the pictorial narrative. Suppose we take the three pictures in the cube comic to be a unit square. Then in the first picture, the point (0.3,0.35) is roughly in the center of the projections of the dark cube, in each of the three pictures. The syntax (7) introduces discourse referents for the three depicted dark cubes in the three pictures, by interleaving the point (0.3,0.35) after each picture, in order to introduce a discourse referent for the dark cube that is depicted in the previous picture.

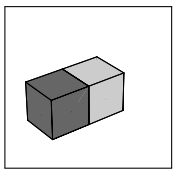
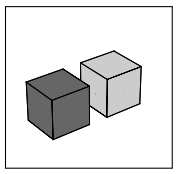
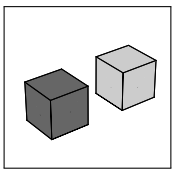
(7)

 (0.3,0.35)  (0.3,0.35)  (0.3,0.35)

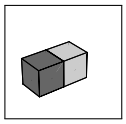
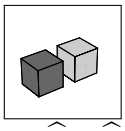
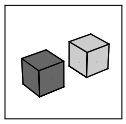
At this level, it is still left open whether the three depicted dark cubes are the same object, or different objects. But introducing discourse referents gives the semantics access to handles for the three depicted dark cubes. In Section 2, the semantic interpretation of a single picture was a relation between worlds and viewpoints. What is the interpretation of a sequence of pictures, with interleaved introductions of discourse referents? A simple answer is that each introduction of discourse referent introduces an additional position in a relation that interprets the picture sequence. Thus while the basic interpretation of a picture is a relation between worlds and viewpoints, the interpretation of (7) is either a five-place relation between a world, a viewpoint for the last picture, and three individuals, or keeping track of the viewpoints for the three pictures, a seven-place relation between a world, three viewpoints, and three individuals.

There are several conventions about how to syntactically reference discourse referents. One option is a recency convention using numerical indices, where 1 references the most recently introduced discourse referent, 2 references the penultimately introduced discourse referent, and so forth. With this, in the context of (7), the formula 1=2 expresses identity between the dark cube in the third panel and the dark cube in the second panel. Similarly, 1=3 expresses identity between the dark cube in the third panel and the dark cube in the first panel. With this, (8) is the syntax that expresses the reading of the cube comic where the dark cubes depicted in the three panels are

identical.

(8)  (0.3,0.35)  (0.3,0.35)  (0.3,0.35) 1=2 1=3

We might also want to express the supposition that the viewpoint stays constant across the three panels. Using $\widehat{1}$ for the viewpoint for the most recent picture, $\widehat{2}$ for the viewpoint for the next to last picture, and so forth, the constancy of viewpoint is expressed by identities such as $\widehat{1} = \widehat{2}$. Introducing discourse referents also for the light cube, the syntax for the co-referential understanding of the cube comic is (9).

(9)  (0.3,0.35) (0.6,0.5)  (0.3,0.35)(0.65,0.5)  (0.3,0.35)(0.75,0.6)

$2 = 6 \quad 2 = 4 \quad 1 = 3 \quad 1 = 5 \quad \widehat{1} = \widehat{2} \quad \widehat{1} = \widehat{3}$

An alternative representation for pictorial narratives with indexing uses the box notation of discourse representation theory (DRT) and segmented discourse representation theory (SDRT). A representation corresponding directly to (9) has six discourse referents for individuals, three for viewpoints, three for worlds, and equalities between them, see (10). The DRT notation reifies discourse referents, rather than using the recency convention. And in the version here, it writes the projection relation into the representation. Application of DRT representations in pictorial semantics are found in Abusch (2012), Bimpikou (2018), Maier and Bimpikou (2019), Maier (2019), Abusch and Rooth (2023a), Schlöder and Altshuler (2023).

(10)

$x_1\ x_2\ x_3\ x_4\ x_5\ x_6$
$v_1\ v_2\ v_3\ w_1\ w_2\ w_3$
$x_5 = x_3\ x_5 = x_1$
$x_6 = x_4\ x_6 = x_2$
$v_3 = v_2\ v_3 = v_1$
$\pi(w_1, v_1) =$ 
$\pi(w_2, v_2) =$ 
$\pi(w_3, v_3) =$ 

4. Multi-modality

Multimodal artifacts are ones that convey information in more than one medium—prominently, a pictorial medium and a linguistic one. They include silent films with intertitles, comics with speech and thought bubbles, children's picturebooks, and pictures with linguistic annotations. The analytic challenge they pose is to characterize how information from two media is integrated into an informational whole. For example, the image in Figure 4 is a redrawn page from a children's

Figure 3: Page from *Gaspard and Lisa's Christmas Surprise* illustrating cross-modal indexing (Gutman, 2002). Linguistic and pictorial parts convey information about the same individuals and events.

picturebook. Assuming a semantics for pictures, and an independent semantics for the English language, we can get to a semantic value for the pictorial part of the page (the pictorial semantic value) and to a semantic value for the linguistic part (the linguistic semantic value). Clearly though, there are relations between the two of them. For one, the individuals depicted in the picture are also designated with nominal phrases in the language. For another, events that are mentioned in the language are often depicted or partially depicted in the image. This is the phenomenon of cross-modal indexing. In natural language syntax and semantics, it is common to work with a co-indexing relation represented as in (11), or alternatively using discourse referents, which are variable-like objects occurring in a representation of discourse semantics and pragmatics. Somehow we would like to apply indices or discourse referents in a way that the indexing relation can cross depictions and mentions in different media. And more generally, we would like to have a framework which allows information from different media to be integrated.

(11)a. Amy$_2$ presented her$_2$ proposal.
b. Every girl$_3$ presented her$_3$ proposal.

A simple case of the problem is provided by tagged pictures, which are pictures with superimposed words that describe depicted objects (Greenberg, 2019). In Figure 4 we see a picture (call it p) of a New England town, with superimposed annotations "Village Green", "Town Hall", and "Congregational Church". According to the geometric analysis from Section 2, the information provided by picture p is a relation that holds between a world w and a viewpoint v iff $\pi(w,v) = p$. The annotations can be analyzed as providing predications such as $\mathbf{townhall}(w,x_1)$, understood as "x_1 is a town hall in the described world w". This linguistic information needs to be combined with

Figure 4: A picture with linguistic annotations, with superimposed points in blue marking introductions of pictorial discourse referents. A linearized version is interpreted compositionally, in a way that captures cross-modal indexing, and combines pictorial and linguistic information conjunctively.
Image ©ⓘMats Rooth. Abusch and Rooth (2020-2026).

the pictorial information in a special way: it's not just that a world w has to contain an object x_1 which is a town hall. Rather, x_1 has to be a town hall that looks like the building depicted in the right part of the picture.

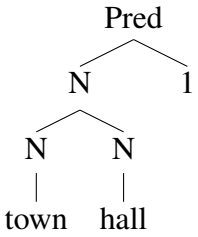
Greenberg (2019) suggested adding points to the syntax of tagged pictures, which serve to link pictorial information with linguistic information. Each point is near in the picture plane to the location of the corresponding annotation. The points are shown in blue in Figure 4. These points can be identified with the geometric points that introduce discourse referents for depicted individuals in the analysis from Section 3. Suppose Figure 4 is partially rewritten as in (12). The first element is the picture; the second element is a geometric point a, which is the position of the rightmost blue dot in Figure 4. It introduces a discourse referent for the town hall as depicted. The third element is the English phrase $[_{\text{N}}$ town hall], combined with the index 1, which in the semantics from Section 3 picks out the most recently introduced discourse referent.[3] This linear notation combines pictorial information (the picture at the start) with English syntax (the small syntactic tree at the end), plus syntax related to pictorial discourse referents (the geometric point in the middle). The result is that a witness for the linguistic information has to be a town hall in the described world w, and has to look like the right part of the picture in world w from viewpoint v. This combines pictorial information with linguistic information, via a mechanism that interprets formulas with pictorial and linguistic parts.[4]

[3]To complete the formula, predications involving the village green and the Congregational church should be added to the right. Each of them is similar to the second and third elements in (12).

[4]Above it was mentioned that using points to introduce discourse referents resulted in ambiguity. If the roof part of the town hall, as well as the town hall, are individuals in the model, a witness for the discourse referent introduced by

(12) 

a



Generalizing, expressions like (12) are formulas of a language that mixes pictures with phrases in the logical form of English syntax. Such formulas can express indexing across pictorial and linguistic modalities. And because the interpretation of the linear notation is conjunctive, the picture and the annotations constrain the same world, and integrate pictorial and linguistic information in the possible worlds framework.

5. Stylized pictures

The geometric projection procedure from Section 2 and the associated semantics is adapted to interpreting images of regular polyhedra like the ones in Section 2, the photo-realistic image of a cat in a box in (13a), and passages of film. It is not on the face of it adapted to interpreting images like the stylized line drawing of a cat in a box in (13b). Equally, it is not adapted to interpreting artistic images like the art nouveau (13c) or expressionist (13bd). Or more cautiously, to use the projection semantics for (13b), (13b), and (13c), one would have to identify projection functions that map worlds with real cats in real boxes to these images. The reason is that the semantic-value function is supposed to be obtained by inverting projection.

(13) a. 

b. 

c. 

d. 

Images ⓒⓘMats Rooth. Abusch and Rooth (2020-2026)

The issue of how artistic style factors into the geometric pictorial semantics has not been faced squarely enough. Two exceptions are Abusch (2012), and in greater depth Maier (2023). Abusch referenced algorithms tuned from data that modify the style of an image to a specified artistic style, call it s (Reynolds, 2002). If stylization is expressed by a function f^s that maps realistic pictures to stylized ones, then the basic projection function π can be composed with stylization to form a function $f^s \circ \pi$ that maps pairs of a world and a viewpoint to a stylized picture. This function can be inverted as before, to obtain semantic values for stylized pictures. This approach is taken by

a could be either the roof or the town hall. The indeterminacy is removed by the linguistic phrase in formula (12).

both Abusch and Maier.[5]

It is unrealistic, though, to assume that the result of stylizing a given realistic picture in a given artistic style is unique. If we ask an artist to redraw the picture of the cat in the box in expressionist style, the result is likely to be different on different occasions. This is paralleled in AI models that are capable of mapping a realistic image to a given artistic style – these come with random seeds and temperature parameters that affect the result. So it is more realistic to assume a stylization relation r^s that maps a given realistic image to multiple stylized counterparts. The combination of a projection function π with a stylization relation r^s can still be inverted as in (14) to obtain semantic values. In words, the semantic value of a stylized picture q is the relation that holds between a world w and a viewpoint v iff q can be obtained by stylizing in style s the value of the projection function π on w and v.

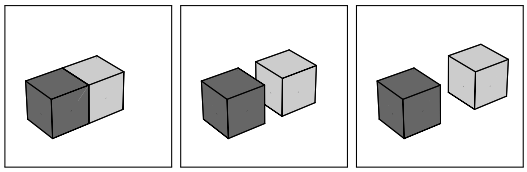
(14) $[\![q]\!] = \{\langle w,v\rangle | \exists p.\pi(w,v) = p \wedge r^s(p,q)\}$

Maier (2023) argued that this method should be applied even to emojis, which are an extreme case of stylization. On this, see Chapter XX in this volume.

6. Temporal progression

There is a temporal aspect to the three-panel sequence repeated in (15). The sequence contributes information about some cubes being in certain configurations, so that they look a certain way from certain viewpoints. Indexing adds the information that the dark cubes depicted in the three panels are the same object, and likewise for the light cubes. Beyond this, it is understood that the time when the dark cubes are one half cube dimension apart follows the time when they are in contact. And the time when the two cubes are one unit apart follows the time when they are half a unit apart.

(15)



In a similar way, when a film is assembled from successive shots, there is a default interpretation of temporal progression. A modern series episode has on the order of a thousand shots, and a modern film one thousand to four thousand shots. Successive shots are typically interpreted with temporal succession, often close temporal succession. There are other interpretations, such as portraying the memories of a character, or flashing back to earlier events in the described world. On this, see the next section.

Indexing and temporal progression are aspects of construing a picture sequence as a narrative or story, where characters recur, and where events and appearances are described in a way where temporal order in the described world is homomorphic to linear ordering of panels or shots, or equivalently to the order in which panels are read or the film is viewed. Remarkably, even purely pictorial narratives, such as graphic novels with no language (no captions, speech bubbles, or though bubbles) can for a human experiencer cohere into a story, including not just attribution of

[5]Or in some passages, Maier seems to contemplate building stylization into the basic projection function.

identity relations and temporal relations, but also causal connections between events, attribution of beliefs and motives to characters, a sense of identification with a protagonist, and so forth.[6]

A way of theorizing about temporal progression is to posit that pictorial narratives are built up structurally. Suppose φ is a structured representation of the first two panels in (15), which includes the first two panels, together with additional information about indexing. The structure φ is to be extended with the third panel, call it c. The extended structure has two parts, one of which is φ, and the second of which is the added panel. We represent the result as a typed data structure, which is displayed as (16). The type label *Narration* is adopted from literature on discourse structure in natural language (Asher and Lascarides, 2003). The intent is that the data type *Narration* marks a discourse structure that is interpreted with temporal progression.[7]

(16) $$\begin{bmatrix} & \text{Narration:} & \varphi \\ \text{Narration} & \text{Picture:} & c \end{bmatrix}$$

There is a worry that hypothesizing a syntax like this entails commitment to something for which the evidence is bound to be indirect. But turning this around, abstract syntax for pictorial narratives is the minimal assumption required to support compositional semantic interpretation. In this way the commitment involved in positing syntax is minimal, or nearly so. Plainly, motivation must come from what it allows one to do in semantic and pragmatic interpretation.

Abusch (2014) looked at phenomena of temporal relations in comics, and compared them to temporal phenomena in natural language. A basic issue is defining the relation of temporal progression on the semantic side. Let t_i and t_{i+1} be two time points, understood to be times when a single world looks like successive panels p_i and p_{i+1} in a comic. (17) describes two possible semantic definitions of temporal progression. The first is strict temporal succession: t_{i+1} strictly follows t_i. The second interpretation of temporal non-regression allows for t_{i+1} and t_i to be the same, in addition to allowing t_{i+1} to follow t_i.

(17) $t_i < t_{i+1}$ temporal succession
$t_i \leq t_{i+1}$ temporal non-regression

There are a couple of varieties of panel and shot sequences where there is little feeling of temporal succession.[8] In film, an establishing shot—often a wide shot—is used to convey information about the location for events depicted in subsequent shots (Sargent, 1913). In graphic novels, there are establishing panels with the same function. Often they are comparatively large, see the example in Figure 6. Often, it seems indeterminate whether the temporal relation implied between an establishing shot or panel and the subsequent one is temporal succession as in (17a) , or the weaker (17b). Bordwell et al. (2020) state explicitly that temporal overlap is an option: "Editing usually presents a series of shots that are temporally continuous, but sometimes the editing repeats part or all of an action, so that time seems to overlap." For film, it is necessary to refer to time intervals, and the weak interpretation that allows for overlap or succession is (17), using notation from Allen's temporal logic (Allen, 1983).

[6]See for instance Masashi Tanaka's *Gon* stories about a small dinosaur in the world of modern animals (Tanaka, 2011), and Shaun Tan's *The Arrival*, telling the story of a man emigrating to a strange city (Tan, 2006).

[7]Here we mean temporal progression, not just default temporal progression. A sequence with a different interpretation has another type.

[8]The phrasing "little feeling of temporal succession" is from McCloud's book-length graphic study of the structure of comics and information in them McCloud (1993).

Figure 5: Establishing panel with a bathhouse below a mountain, followed by four sequenced panels depicting sequenced events inside. The reading order is right to left. Tezuka (1972).

(18) $i_k \circ i_{k+1} \vee i_k < i_{k+1}$ overlap or succession between temporal intervals

This discussion brings up the issue of how fine-grained the vocabulary of narrative structures of visual narratives is. Are an establishing panel or shot and the subsequent one combined under *Narration*, or is there more specific structure such as *Establishing*? Manuals of screenwriting and film editing, and empirically oriented literature on film structure have developed fine typologies, without really needing to answer the question whether a particular transition is an instance of a fine category, or a superordinate one. Our impulse is that, in order to connect with empirical phenomena, as a research strategy one should work with fine vocabularies.

In theorization about natural language, on top of temporal relations, discussion of aspectual structure is prominent (Dowty, 1979). The most important aspectual distinction is the opposition between eventives and statives. "All languages recognize this aspectual contrast and many grammaticalize it in various ways" (Bittner, 2014). There are many linguistic tests that diagnose the distinction. In English, stative sentences can be modified by *still* with a temporal sense (19a,b), and cannot be used in the progressive (19c). Abusch (2014) considered whether there is a distinction between stative and non-stative panels in pictorial narratives. Some phenomena are suggestive of a stative status for some panels. The information in establishing panels can be viewed as stative, in that there is an implication that the geometric information in the establishing panel extends over subsequent ones. In the example in Figure 6, there is an implication that the geometry of a mountain above a guesthouse is maintained over the events depicted in the subsequent panels. This can be held to parallel the situation in natural language, where in (20), the state of coldness and darkness is understood to overlap the events of bumping, grunting, awaking and grabbing.

(19)a. Your glasses are still on the counter. (A stative predicate with temporal *still*.)

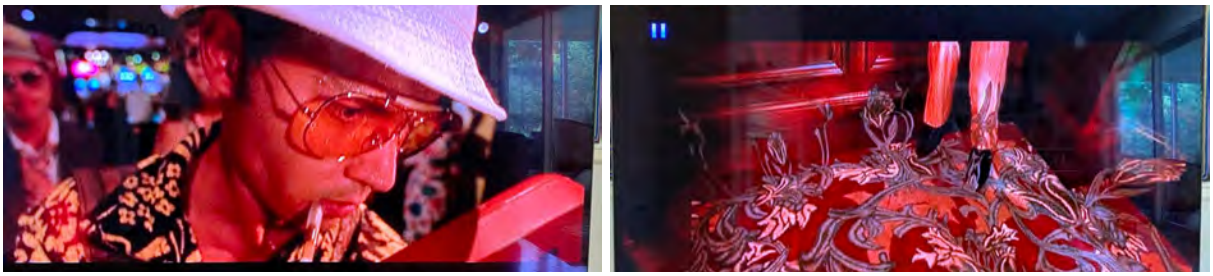


Figure 6: Duke looks down, and hallucinating, sees paisley patterns creep up the walls and the legs of the man in front of him. *Fear and Loathing in Las Vegas*, 1998. Johnny Depp (Raoul Duke).

b. I still placed your glasses on the counter. (A non-stative predicate, but not a temporal sense for *still*.)
c. *Your glasses are being on the counter.

(20) It was cold and dark in Jack's tent. A black bear bumped against the back pole and grunted. Jack awoke and grabbed his jackknife from his backpack.

Literature on natural language holds that at least in some cases, an interpretation of overlap depends on the aspectual status of a clause as stative. But Abusch (2014) argued that for pictures, there is no semantic distinction between stative and non-stative information. A basic reason is that geometric information, such as the book being on the table, is characteristically stative in natural language. And the projection procedure from Section 2 is geometric. Abusch gives additional arguments and concludes that pictorial information is always semantically stative. If there are no aspectual distinctions in the component parts of pictorial narratives, then it is impossible to make temporal succession sensitive to aspectual status.

7. Embedding

Figure 6 shows some frames from a scene in the film *Fear and Loathing in Las Vegas*. The protagonist Duke is described in the film as being on hallucinatory drugs. The shots show paisley patterns creeping up the walls and the legs of the man in front of him. The base worlds described by the narrative probably never look like these shots, because the shots show what Duke is hallucinating, not how things look while he is hallucinating. It is natural to analyze such shots as being embedded—in this case under an operator, call it *Hallucinate*, which introduces a passage that depicts the hallucinations of a character. This connects with linguistic examples. In (21), there is overt embedding under the verbs *see* and *hallucinate*. In (22), there is no overt embedding, but such examples can be analyzed as involving hidden embedding, either in the construction called free indirect discourse (Eckardt, 2014), or in the construction called protagonist projection (Abrusán, 2021). The terminology of embedding is also current in narrative theory (Nelles, 2002, 2010).

(21)a. Duke saw paisley patterns creep up the walls and his legs.
b. Duke hallucinated paisley patterns creeping up the walls and his legs.

(22) Duke looked down. Paisley patterns were creeping up the walls and his legs.

Figure 7: Mary Holmes makes false statements to the DA and a police detective. The third image is from an embedded sequence rendering what Holmes says, while the second and fourth images are from extensional shots. *The Goose Woman* (1925). Mary Holmes (Louise Dresser), Amos Ethridge (Marc McDermott), District Attorney Vogel (Gustav von Seyffertitz), Detective Kelly (George Nichols).

In the silent film *the Goose Woman*, an elderly woman gives testimony to the DA and a police detective. Part of what she says is rendered in intertitles, but an extended part is rendered in an embedded passage of about twenty-five short shots (Figure 7). It is inferred that the woman is lying—a base world described by the film does not look like the embedded shots.[9]

Against the background of the geometric semantics for pictures and film shots that was outlined in Section 2, the signal property of shots like the ones in Figures 6 and 7 is that worlds described by the film do not have to ever look like the shot. Instead, in Figure 6 it is understood that Duke is hallucinating, and what the agent hallucinates is described by the shot. The entailment that worlds described by the film look like the shot is blocked. Constructions with this property are (in terminology from natural language semantics and the philosophy of language) non-extensional constructions. In the setup from Section 6, panel sequencing and shot sequencing are governed by *Narration*, which expresses temporal progression (or alternatively, temporal non-regression), but maintains the entailment that a described world looks like the comic panels or film shots that are combined by *Narration*. In this conception, the discourse relation of *Narration* combines panels or shots, but it does so extensionally, so that a sequence of panels or a sequence of shots are construed as providing constraints on the same world.

[9]Another character separately makes statements that are understood as true, and are also described by an embedded shot. Since the statements of the two characters are incompatible, the embedded passages cannot both be extensional. See the discussion of *The Goose Woman* in Turim (1989), in the section *Flashbacks rendering verbal narration visual*.

The above leads to the theoretical move of postulating additional discourse structures such as *Hallucinate* that embed panels or shots, but do so non-extensionally. Usually, in current film and series, non-extensional embedding is not marked in any special way. By default, units are combined by *Narration*, but creators can intend and viewers can infer other non-extensional relations such as *Hallucinate*. Luchoomun (2012) gives an empirical typology of non-extensional constructions in film, and Turim (1989) describes the evolution of non-extensional constructions.

A syntax for non-extensional embedding has to keep track of information about agents. When the embedded passage describes the hallucination, statement, or memory of an agent, interpretation needs to make reference to the identity of that experiencing agent, in order to attribute the information to them. This is accomplished by making a discourse referent for the agent an argument of the embedding predicate—this is the index k in (23a) and (23b). Second, and more subtly, the embedded passage commonly depicts the same agent, or a counterpart of that agent. For instance in Figure 7, the embedded passage depicts Mary Holmes, together with some other characters. Thus the passage depicts an individual that the agent informationally takes to be herself. This individual in the embedded passage is picked out with another index, which is the index j (23a) and (23b). So k is the index of the experiencing agent in the outside discourse, while j is the index in the embedded one. In this notation, an equality $k = j$ is not included, because the structure inherently interprets j as a counterpart of k. As exemplified in (23b), different embedding interpretations such as saying use a different parent label, in order to trigger interpretation.

(23) a. Hallucinate
IShot j k

b. Say
IShot j k

c. Hallucinate
IShot k

The embedded shot in *Fear and Loathing* actually does not show Duke or a counterpart of Duke. Instead the viewpoint for the embedded shot is the visual perspective of the Duke. This is a subjective point of view passage, where the film shot shows what the agent sees or hallucinates. Abusch and Rooth (2017) and Abusch and Rooth (2023b) analyzed this construction in comics, with some reference also to film. They suggested a notation equivalent to (23c), where there is an index k for the external agent, but the internal agent is implicit, because that agent can be recovered as an agent who has the same visual viewpoint as the shot.[10] Often such "free perception" shots are set up by preceding extensional shots that show the agent looking or glancing. The terminology of free perception also applies to examples like (24) in natural language, where the embedding is free in that there is no overt embedding of the main clause.[11]

(24) When I looked up a guy with a metal detector was walking toward me.

8. Discussion

This article has reviewed the possible worlds approach to the semantics of image sequences. It is often pointed out that possible worlds semantics for language does not say very much about the semantics of basic lexical items; rather, it is largely concerned with how, assuming semantics for

[10]Rooth and Abusch (2026) is concerned with the visually non-subjective construction in film, and use the syntax (23a), with an index also for the internal agent.

[11]The example is from the story "Ghosts" by Brian Hart. It is implied that the first person protagonist saw a guy with a metal detector walking toward them. An early discussion of examples like this is Brinton (1980).

basic words, the semantics of complex phrases is composed. In contrast, possible worlds semantics for pictorial materials is notably concrete and precise, because of the basis in the geometry of projection.

Something that comes up in several ways is that elements of the abstract syntax that are significant for interpretation, such as introduction of indices, co-indexing, and embedding are not marked in an overt way in comics and film. In natural language syntax, there are constraints on indexing that are typically conceived of as syntactic, and there certainly is embedding that is marked in phrase structure, and as a result in the sequencing of morphemes. It is perhaps remarkable that pictorial narratives can be understood by readers and viewers with, compared to natural language, relatively little evidence for the syntax which determines semantic interpretation, according to the account reviewed here.

We think the most interesting aspect of the theorization reviewed here is that it analyzes language and pictorial materials using a single model of information content. This turns out to almost automatically solve the problem of how to integrate information in multi-modal materials, including allowing for a formalization of cross-modal indexing.

References

Abrusán, M. (2021). The spectrum of perspective shift: protagonist projection versus free indirect discourse. *Linguistics and Philosophy 44*(4), 839–873.

Abusch, D. (2012). Applying discourse semantics and pragmatics to co-reference in picture sequences. In *Proceedings of Sinn und Bedeutung 17*.

Abusch, D. (2014). Temporal succession and aspectual type in visual narrative. *The Art and Craft of Semantics: A Festschrift for Irene Heim 1*, 9–29.

Abusch, D. and M. Rooth (2017). The formal semantics of free perception in pictorial narratives. In *Proceedings of 21st Amsterdam Colloquium*.

Abusch, D. and M. Rooth (2020–2026). Supersemantics: An interdisciplinary reader in philosophy linguistics and psychology and comics studies and film theory and artificial intelligence. `https://github.com/MatsRooth/supersemantics`. GitHub repository, ongoing project.

Abusch, D. and M. Rooth (2023a). Parallel and differential contributions from language and image in the discourse representation of picturebooks. In *Proceedings of Sinn und Bedeutung*, Volume 27, pp. 1–18.

Abusch, D. and M. Rooth (2023b). Pictorial free perception. *Linguistics and Philosophy 46*(4), 747–798.

Alberti, L. (1435). *Della Pittura*.

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM 26*(11), 832–843.

Asher, N. and A. Lascarides (2003). *Logics of Conversation*. Cambridge University Press.

Bimpikou, S. (2018). Perspective blending in graphic media. *ESSLLI 2018 Student Session*, 245–258.

Bittner, M. (2014). *Temporality: Universals and variation*. John Wiley & Sons.

Bordwell, D., K. Thompson, and J. Smith (2020). *Film Art: An Introduction* (12 ed.). New York, NY: McGraw-Hill Education.

Brinton, L. (1980). 'Represented perception': A study in narrative style. *Poetics 9*(4), 363–381.

Cumming, S., G. Greenberg, and R. Kelly (2013, August 25). Chess case. Video, Vimeo. Accessed: 2026-02-16.

Cumming, S., G. Greenberg, and R. Kelly (2017). Conventions of viewpoint coherence in film. *Philosopher's Imprint 17*(1).

Dekker, P. (1994). Predicate logic with anaphora. In *Proceedings from Semantics and Linguistic Theory*, Volume 4, pp. 79–95.

Dekker, P. J. (2012). *Dynamic Semantics*, Volume 91. Springer Science & Business Media.

Dowty, D. R. (1979). *Word meaning and Montague grammar: The semantics of verbs and times in generative semantics and in Montague's PTQ*, Volume 7. Springer Science & Business Media.

Dürer, A. (1525). *Underweysung der Messung mit dem Zirckel und Richtscheyt*. Nuremberg, Germany: Hieronymus Andreae Formschneider.

Dürer, A. (1917). The draughtsman of the lute. Woodcut print, sheet: 13 x 18.2 cm. Object Number 17.37.313; Drawings and Prints. Metropolitan Museum of Art, New York.

Eckardt, R. (2014). *The Semantics of Free Indirect Discourse: How texts allow us to mind-read and eavesdrop*. Brill.

Greenberg, G. (2019). Tagging: Semantics at the iconic/symbolic interface. In *Proceedings of the 22nd Amsterdam Colloquium*.

Greenberg, G. J. (2011). *The Semiotic Spectrum*. Ph. D. thesis, Rutgers University.

Gutman, A. (2002). *Gaspard and Lisa's Christmas Surprise*. New York: Alfred A. Knopf.

Heim, I. (1982). *The Semantics of Definite and Indefinite Noun Phrases*. Ph. D. thesis, University of Massachusetts, Amherst.

Kamp, H. (1981). A theory of truth and semantic representation. In *Proceedings of the 3rd Amsterdam Colloquium*, pp. 277–322.

Lempitsky, V., P. Kohli, C. Rother, and T. Sharp (2009). Image segmentation with a bounding box prior. In *2009 IEEE 12th international conference on computer vision*, pp. 277–284. IEEE.

Lewis, D. (1972). General semantics. In *Semantics of natural language*, pp. 169–218. Springer.

Lewis, D. (1975). Adverbs of quantification. In *Formal Semantics of Natural Language*, pp. 178–188. Cambridge University Press.

Luchoomun, L. (2012). *Mental images in cinema: Flashback, imagined voices, fantasy, dream, hallucination and madness in film*. Ph. D. thesis, Roehampton University.

Maier, E. (2019). Picturing words: the semantics of speech balloons. In *Proceedings of 22nd Amsterdam Colloquium*.

Maier, E. (2023). Emojis as pictures. *Ergo 10*(12), 317–355.

Maier, E. and S. Bimpikou (2019). Shifting perspectives in pictorial narratives. In *Proceedings of Sinn und Bedeutung 23*, pp. 91–105. University of Konstanz.

McCloud, S. (1993). *Understanding comics: The invisible art*. Kitchen Sink Press.

Montague, R. (1973). The proper treatment of quantification in ordinary english. In *Approaches to natural language: Proceedings of the 1970 Stanford workshop on grammar and semantics*, pp. 221–242. Springer.

Montague, R. et al. (1970). Universal grammar. *1974 222*, 246.

Nelles, W. (2002). *Stories Within Stories: Narrative Levels and Embedded Narrative*. Columbus: Ohio State University Press.

Nelles, W. (2010). Embedding. In D. Herman, M. Jahn, and M.-L. Ryan (Eds.), *Routledge Encyclopedia of Narrative Theory*. Routledge.

Rabern, B. (2016). The history of the use of ⟦.⟧-notation in natural language semantics. *Semantics*

and Pragmatics 9(12), 12.

Reynolds, C. (2002). Stylized depiction in computer graphics non-photorealistic, painterly and 'toon rendering. *An annotated survey of online resources www. red3d. com/cwr/npr*.

Rooth, M. and D. Abusch (2026). An existential semantics for over-informative attitudinal embedding in film. In *Proceedings of Sinn und Bedeutung*, Volume 30. Universität Konstanz.

Sargent, E. W. (1913). *The Technique of the Photoplay*. New York: The Moving Picture World.

Schlöder, J. J. and D. Altshuler (2023). Super pragmatics of (linguistic-) pictorial discourse. *Linguistics and Philosophy 46*(4), 693–746.

Tan, S. (2006). *The Arrival*. Hodder Children's Books.

Tanaka, M. (2011). *Gon, Vol. 1: Episodes 1–4*. New York: Kodansha Comics. Originally published 1992–1994 by Kodansha Limited (Tokyo).

Tezuka, O. (1972). *Kirihito Sanka*. COM Comics Zōkan. Mushi Pro Shōji. 2 vols. (March–April 1972). Original Japanese edition.

Turim, M. (1989). *Flashbacks in Film: Memory & History*. Routledge.

Vasari, G. (1568). Vita di Filippo Brunelleschi. In *Le Vite de' più eccellenti Pittori, Scultori e Architettori* (2 ed.). Florence: Giunti. Second enlarged edition (1568).