Pictorial language and linguistics*

Emar Maier
University of Groningen

Abstract A language is a system of signs used for communication, and linguists are tasked with, among other things, uncovering the syntax and semantics of such systems. In this paper I explore to what extent pictures fit this characterization of a language and hence would fall within the domain of linguistics. I conclude that at the very least there are well-defined systems of depiction for which we can give a precise semantics, in a possible worlds framework reminiscent of what linguists are used to. The only real difference is that pictorial propositions are not derived via the linguistically familiar lexicon and recursive composition rules, but via geometric projection. I then show how sequences of pictures, like sequences of utterances, can be used to form coherent discourses. I explore in some detail extensions of the formal discourse semantics of Segmented Discourse Representation Theory to deal with comics, i.e. storytelling through picture sequences. I focus specifically on the representation of events and the integration of both symbolic and iconic elements in comics.

Keywords: Super Linguistics, picture semantics, geometric projection, viewpoint, visual narrative, discourse structure, coherence

Contents

* Second draft of a paper written for the *Oxford Handbook of Philosophy of Linguistics*, edited by Gabe Dupré, Ryan Nefdt, and Kate Stanton. I thank Gabe and Kate in particular for their invitation, patience, and valuable comments on the first draft.

1 Varieties of meaning

Paraphrasing Saussure (1916) language may be defined as a system of signs used for communication. In this definition, *communication* is an exchange of information, and a *sign* is an object or event that stands for something beyond itself, like the written inscription 'cat' stands for a certain species of animal. The key ingredient of this definition, that sets language apart from other kinds of signals we find throughout nature, is the notion of a *system*, which is typically cashed out in terms of a syntax (a set of rules for systematically generating well-formed signs), coupled with a semantics (a set of rules matching all well-formed signs with what they stand for).

In the wake of Frege (1892), Chomsky (1957) and Montague (1973), theoretical linguists have concerned themselves with constructing recursive grammars with matching compositional semantic interpretation rules that together generate predictions that are supposed to match native speakers' intuitions about grammaticality and acceptability for a vast range of languages, in the spoken or signed modality. But there are many different "systems of signs used for communication" beyond Dutch or ASL. Traffic signs, to take a very simple example, clearly fit the Saussurian definition of a language. They are *about* traffic situations and road usage; they are used to *communicate* what actions are allowed or prohibited; and they are *systematic*, i.e., they have a clear syntax (some combinations of colored shapes on a metal plate count as well-formed traffic signs, others do not); and their meanings systematically depend on certain aspects of their forms (e.g., a red circle always means that the action depicted inside is prohibited, while a blue background always describes what situation you're in).

If we now look a little closer at the way traffic signs convey their meanings we may notice that there is something different about the way a sign like (1a) and one like (1b) express that you're on a bike lane.

(1) a. 

b. 

Both are signs in a Saussurian sign system of traffic signs, describing that there's a bike lane there. But (1a) *describes* the biking aspect of the sign's meaning verbally, where (1b) *depicts* it. The Dutch word 'fietspad' in (1a) refers by virtue of a rather arbitrary connection between inscriptions and meanings, one that a foreign visitor might have to look up somewhere before she can interpret it. The sign in (1b) is considerably more transparent, in the sense that most foreign visitors will immediately understand it has something to do with bicycling. In semiotic terminology, the word 'fietspad' in (1a) is a symbolic representation, while the bicycle picture in (1b) is iconic.

Linguists tend to focus on what I call verbal languages, i.e., sign systems that allow building meaningful structures out of smaller meaningful elements, viz. words (or morphemes), in either the spoken, written or signed modality. Pictorial signs like drawings or photographs are not usually considered verbal in this sense. In fact, it is far less clear than for traffic signs that such signs can plausibly be captured under the Saussurian definition of a language, that is presupposed in contemporary linguistics. Is there really a systematic syntax and rule-based 'semantics of sketching'? And if so, can we formalize and study such systems at anywhere near the level of precision as the semantics of English?

In this chapter I suggest a positive answer to these questions, paving the way for a linguistics of pictures. Concretely, after providing some background on semiotics and linguistics (Section 2) I spell out what a syntax and semantics of pictures might look like (Section 3). I will make use of the toolkit of formal semantics, so let me already briefly elaborate why we should want to use that kind of framework. In addition to the general, theoretical selling points of formal semantics – like the built-in account of intentionality that we get by cashing out meaning as a relation between language and world, rather than as a relation between words and mental constructs – I will demonstrate in the second half of the paper how a formal semantics of pictures opens up a vast, extended toolkit for studying meaning and interpretation, including powerful theories of discourse structure and pragmatics. I show how we can use this to study the meaning and use of various pictorial and multimodal artifacts that we engage with on a daily basis to satisfy our need for communication and storytelling. I will explore one concrete application, viz. the use of linguistic discourse coherence theory to better understand pictorial storytelling in comics (Sections 4–5).

2 Background: Symbolic and iconic sign systems

Peirce's (1868) distinction between icon and symbol has been applied and extended in a range of humanities and social sciences fields concerned with meaning and communication. Peirce's own characterization is in terms of resemblance vs arbitrariness: an iconic sign resembles its referent (a picture or statue of Napoleon refers to Napoleon because they look similar) while a symbolic sign relies on an arbitrary convention to get its meaning ('regen' in Dutch and 'pluie' in French both mean rain, for no particular reason other than that that's what the respective lexicons say). In addition, Peirce also introduces the index, a type of sign that refers by virtue of a factual or causal connection (smoke means fire, clouds mean rain). This third category corresponds roughly to what Grice (1957) calls 'natural meaning', i.e., meaning that exists independent of any human intentions to communicate and hence seems prima facie less suitable for linguistic analysis.

Natural languages like English are often assumed to be largely symbolic. This is most obvious at the ground level, where words pick out their meanings. Evidence for the defining arbitrariness of word meaning is the enormous cross-linguistic variety of word inventories and meanings, and the observation that acquiring new languages requires learning this new vocabulary. At the higher levels, meaning depends more systematically on grammatical structure and context, but still there is evidence of apparent arbitrariness, evidenced by crosslinguistic variation and an apparent need for explicit learning of the inventories and usage patterns of, for instance, speech reporting strategies, evidentiality, mood, tense, aspect and honorific marking, the expression of questions and commands, etc.

By contrast, photos, drawings, statues, and video games are usually considered iconic. Peirce originally coined the term 'Likeness' for what he later called icons, defining them as signs that refer by virtue of a resemblance, i.e., the sharing of a certain salient quality. A statue of Napoleon shares a salient quality (3D shape) with the actual Napoleon that it refers to. For photos or drawings this is a bit more complicated, but we might still say that a photo of my cat shares some key spatial properties with the actual animal. In other words, the mechanics of photography instantiates a 3D-to-2D transformation that preserves relevant form aspects. Realistic landscape painting would likewise instantiate a form-preserving transformation, and even maps or stick figure drawings preserve some relevant form aspects (while presumably modifying, approximating, ignoring, or exaggerating other form aspects). If we want a fully general definition of iconicity that includes even Venn diagrams (Shin et al. 2018), or ideophones (Dingemanse 2012), or even fuel gauge dials (Greenberg 2022), we need to go beyond the intuitive notion of a form-preserving transformation, but for pictures (photos, sketches, technical drawings, paintings, comics, film) it seems like a solid starting point.

In the next section I zoom in on pictorial representation and argue that it makes sense to talk of pictorial languages and to borrow some formal tools and terminology of theoretical linguistics to investigate them. In the next section, while looking for pictorial analogues of syntax and semantics, we make the intuitive characterization of iconicity as involving form-preserving transformations more precise by equating those transformations in the case of pictures with 3D-to-2D geometric projection functions.

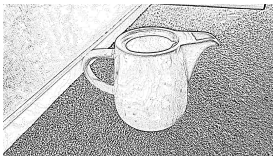
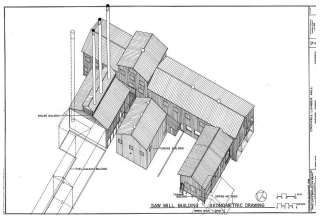
3 Towards a linguistics of pictures

In this section I explore to what extent pictorial representation can be made to fit under the Saussurian language definition: a language is a system of signs used for communication. I take it as a given that pictures are signs that can be used to communicate (i.e., to transmit information about what the world is like, how one should behave, or what happens in some imaginary storyworld). What's left is to explain their salient communicative functions in terms of some underlying systematicity, as opposed to mere ad hoc or purely natural resemblance relations. Specifically, in this section we're looking for a rule-based syntax and semantics of depiction.

3.1 Pictorial syntax I: grammaticality and decomposition

In a sense, looking for a syntax of pictures seems futile – any combination of lines on a canvas, or pixels on a screen is a picture. Put differently, you can't put some lines or pixel colorings together to create an 'ungrammatical picture', in the way that syntacticians move some words around to create ungrammatical sentences, i.e., strings of sounds or inscriptions that aren't sentences at all. On the other hand, a string of sounds is never ungrammatical as such either, only relative to a specific language. Just as there are many different spoken and signed languages, we must first try and distinguish between different pictorial languages, and it is then only in these individual languages that we can look for a pictorial analogue of (un)grammaticality.

An example of what we will consider a specific pictorial language is that of linear perspective grayscale pixel drawing, henceforth $\mathscr{L}_{linear}$, exemplified in (2a). A different pictorial language would be axonometric line drawings ($\mathscr{L}_{axon}$, (2b)), or curvilinear ('fish-eye') circular colour photography ($\mathscr{L}_{curvi}$, (2c)).

(2) a.  b.  c. 

Given a specific pictorial language (or dialect) we could now look for a rule system that restricts the combination of pictorial elements (e.g., pixels, lines, shapes, drawings of individual objects) into more complex pictures.

Cohn (2013b) goes down this path in his analysis of comics panels. For instance, he distinguishes various 'morphemes' like starry eyes or the idea light bulb that can be combined in certain ways with drawings of faces to create a more complex picture, expressing a complex meaning. As he shows, such morpheme combinations are highly constrained, just as in spoken language. For instance, the idea light bulb, indicating inspiration, is what he calls an upfix; it becomes ungrammatical when placed inside a thought or speech bubble (3b), or when placed next to rather than above the head of a character (3c).

(3) a.  b. *  c. *  (from Cohn 2013b)

In recent work, Greenberg (2022) and Lande (2024) likewise analyse pictures as essentially composed of more primitive elements like pixels and lines, which they also explicitly compare to the lexical elements of verbal languages, being composed according to a rule-based grammar. Similar ideas about syntactic structure and decomposition of pictorial representations have arisen in discussions about the nature of visual mental states (Kosslyn 1980, Lande 2020).

Picture decomposition approaches to pictorial syntax seem particularly well-suited for comics, anime, or emojis, where we see pictures decorated with non-pictorial elements (like speech balloons and motion lines, see Section 5.3 below), and perhaps also for (children's) schematic stick figure drawing styles (Wilson & Wilson 1977). It seems less plausible in the realm of photography or realistic landscape painting. Such pictures don't seem to be created and appreciated more holistically. In so far as we do find some kind of conceptual decomposition in the cognitive processing of a photo this is plausibly just the result of more general principles of the visual system's natural processing of visual inputs involving attention and Gestalt principles, rather than language-like structures imposed by the artist's composition. Unlike with Cohn's examples in (3), there are no categorical constraints leading to

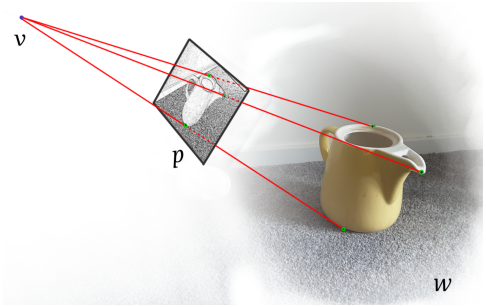
apparent ungrammaticality here. We can draw an extra tree, or an alien space ship, into any Bob Ross painting and it will still be a well-formed, interpretable painting (though the addition might turn it into a very strange, unrealistic, non-truthful, and/or ugly painting). We'll revisit the question of language-like meaning composition at the level of discourse below in Section 4.

3.2 Pictorial syntax II: construction via projection

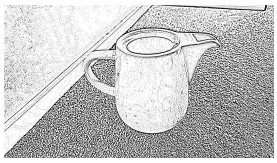
Looking for a proper pictorial analogue of a linguistic grammar we might replace the above characterization of syntax as the study of (un)grammaticality as such, with a more cognitive/communication-theoretic characterization of syntax as the study of (the rule system underlying) the production of proper signs to express a given content (Kamp 2015, Blutner et al. 2006, Hendriks & De Hoop 2001). Departing radically from the generative syntactician's 'autonomy of syntax' (with semantics and phonology as distinct interfaces), this view of language defines syntax as essentially the opposite of semantics – the study of (the rule system underlying) the interpretation of a given sign. On such a production-based conception of syntax we actually do have a pretty robust theoretical understanding of its pictorial analogue. Specifically, thanks to advances in such fields as (computer) vision, visual arts, and geometry we have well-established formal accounts of projection, i.e. the algorithms for turning a 3D scene into a faithful 2D picture of that scene (Hagen 1986, Willats 1997, Greenberg 2013)

Linear projection, for instance, works roughly as follows: (i) choose a viewpoint v (i.e., a unit vector located somewhere in the 3D world representing the line of sight of a (virtual) observer or camera); (ii) assume a so-called picture plane p somewhere perpendicular to that line of sight; (iii) draw so-called projection lines connecting (relevant parts of) objects to be depicted with the viewpoint v; (iv) mark every pixel or point where a projection line crosses the picture plane.

(4) a.



b.



Note that this informal description leaves room for various parameter settings. $\mathscr{L}_{linear}$, as exemplified in (4), uses linear perspective, with a rectangular, flat, white picture plane orthogonal to the viewpoint vector, and projection parameters set to

approximate colors of non-edge surfaces with corresponding grayscale pixel markings. The common technical drawing style $\mathscr{L}_{axon}$, (2)b, by contrast, uses parallel perspective, which relies on projection lines running parallel to each other and to the viewpoint vector. Drawing projection lines in parallel eliminates the characteristic distortions of linear perspective where objects further from the viewpoint appear smaller. In addition $\mathscr{L}_{axon}$ uses projection parameters that ignore all colors and surface structure, merely marking projection lines coming from (sharp) edges of objects. $\mathscr{L}_{curvi}$, (2)c, uses curvilinear perspective, which can be described as involving linear perspective projection onto a spherical surface centered around the viewpoint, followed by a parallel perspective projection of the spherical image onto a tangential flat plane. Further projection parameters of $\mathscr{L}_{curvi}$ have been set to approximate surface colors with a fixed palette of pixel colors.

In sum, projective geometry provides a formal framework for understanding some of the familiar ways of turning a 3D scene into a 2D depiction of that scene. Insofar as we can characterize syntax in communication-theoretic terms, as the study of the mechanisms underlying the production of well-formed signs in a given language, we can view this as a plausible analogue of a grammar in the pictorial domain. Different projection functions then correspond to the different grammars of different languages. Crucially, projection functions differ from familiar linguistic grammars in that they work much more holistically: neither input (scene) nor output (picture) is necessarily chunked up into hierarchically ordered parts at any point in the construction process.

A final remark about picture construction is that outside the realm of photography we frequently depict situations that are not real. As with verbal language, we can lie, or at least mislead, and tell fictional stories. Since projection is literally a function mapping a 3D scene to a 2D picture, our grammar seems to capture only the depictions of actual scenes, as in (unmanipulated) photography. To deal with fiction, lying and misrepresentation we could look to linguistic semantics and appeal to (sets of) possible worlds. A fictional drawing then can be seen, roughly, as the result of projecting in some pictorial language a 3D scene occurring in a non-actual possible world. In a way, the fact that we're bringing in possible worlds does highlight the limits of our pictorial syntax metaphor and shows how we've drifted into the territory of semantics. We will leave it at this and discuss the role of possible worlds in pictorial semantics below.

3.3 Pictorial semantics

Semantics is the study of the relation between signs and their meanings. Let's review what the semantic interpretation of a sentence, arguably the most fundamental unit of verbal communication, looks like.

(5) I see a teapot.

Following a standard compositional semantic approach (Heim & Kratzer 1998), the interpretation of (5) involves, first, the determination of a corresponding syntactic tree structure (Logical Form, LF) which in turn determines a series of functional applications applied to a few primitive lexical entries (e.g., the entry associating 'teapot' with a function that maps possible worlds to the sets of all teapots in those worlds). The end result of the compositional semantic computation will be a possible worlds proposition that captures the intuitive truth conditions (and thus the meaning or informational content) of the sentence, relative to a specific utterance context c (Kaplan 1989).

(6) $[\![(5)]\!]^c = \{w \mid \text{in } w \text{ there's a teapot } x, \text{ and the speaker of } c \text{ sees } x \text{ at the time of } c\}$

A pictorial semantics works rather differently. Crucially, as we've observed above, many purely pictorial signs have no syntactic structure and are thus interpreted holistically as opposed to compositionally. In other words, a semantics for a certain pictorial language, like $\mathscr{L}_{linear}$ or $\mathscr{L}_{curvi}$, will not rely on a structured LF, nor a lexicon of primitive meaningful picture constituents, nor a set of recursively applicable interpretation rules. But that doesn't mean we have to abandon the hope for a semantic rule system. Instead, we go back to the communication-theoretic view of meaning and characterize comprehension (semantics) as the inverse of production (syntax). Concretely, the semantics of a given pictorial language, say $\mathscr{L}_{linear}$, is then a kind of inverse of the geometric projection function that characterizes picture production in $\mathscr{L}_{linear}$(Abusch 2020). Instead of feeding a world and a viewpoint to get a picture, we feed the picture and reconstruct the world and viewpoint:

(7) $[\![$  $]\!] = \left\{ \langle w, v \rangle \;\middle|\; \Pi_{\mathscr{L}_{linear}}(w, v) = \right.$  $\left. \right\}$

The semantics of $\mathscr{L}_{linear}$ thus assigns to each picture a set of world–viewpoint pairs, viz. those that when fed into the projection function of $\mathscr{L}_{linear}$ would generate that picture. If we think of a world–viewpoint pair as a scene, a slice of the world seen from a specific vantage point, then we can informally paraphrase the meaning of a picture as the set of scenes that look like that.

Although high-resolution photos can give lots of information about the shape or color of a scene, perhaps more than we could describe in a thousand words, there are still always infinitely many possible worlds and viewpoints that would give rise to that same picture, if only because the picture cannot tell you what happens behind the viewpoint, or after the photo was taken, or exactly how far away or big things are.

3.4 Comparing verbal and pictorial semantics

The output of our holistic picture semantics applied to the $\mathscr{L}_{linear}$ teapot picture in (7) is quite similar to the output of the compositional English semantics applied to the English teapot sentence in (6). Both consist of an infinite set of alternative possible states of affairs that contain a teapot, or in the case of (7) a roughly teapot-shaped object.[1]

One salient difference is that in (6) we relativize interpretation to a context of utterance and as a result we just get a classical (or 'horizontal') proposition, a set of worlds, while in (7) we take the location of a viewpoint as part of the information conveyed by the picture, leading to a more finegrained '(viewpoint-)centered proposition'. Now, the notion of a viewpoint, the point from which the world is shown in the picture, is closely related to that of a context of utterance, the point at which the utterance is produced. We can bring out this analogy by thinking of the Kaplanian context parameter c slightly more generally as a 'context of creation', which includes parameters for an agent a_c (i.e. the sign's producer, speaker, thinker, viewer, painter etc.), a possible world w_c, a time t_c, and a viewpoint v_c (i.e., the relevant gaze direction of a_c in w_c at t_c). Now note that, just as with verbal utterances, we can define the classical proposition expressed by a picture relative to a given context of production (and a pictorial language):

(8) $[\![$  $]\!]^c = \{ w \mid \Pi_{\mathscr{L}_{linear}}(w, v_c) =$  $\}$

The reason why (8) is not quite as useful as its verbal counterpart in (6), or as the viewpoint-centered version in (7), lies in the default use cases of verbal vs pictorial language. In linguistics, we tend to think of language being spoken in a direct face-to-face interaction between a speaker and a hearer, in which case the interpreter has a pretty good sense of what all the relevant parameters of the context of creation are. The interpretation of a picture, by contrast, typically occurs at a different time and place from its creation, and it is indeed quite rare for the viewer of a painting to know in advance – before interpreting what's on the painting – where in the world a painter was located at the time of painting.

The same separation between producer and consumer actually also holds for many cases of verbal language use, ranging from an answering machine message to a printed book. For such cases Stalnaker (1978) had already introduced the concept of a diagonal proposition, defined as the set of contexts in which the classical proposition expressed by an utterance would have been true (Zimmermann 1991):

1 Strictly speaking, the object depicted in (7) need not be teapot-shaped at all, as we could be dealing with a picture of a picture of a teapot (see Kulvicki 2013).

(9) $\backslash\backslash \varphi \backslash\backslash = \{ c \mid w_c \in [\![\varphi]\!]^c \}$

Diagonalization effectively makes the context a part of the content expressed, so, for instance, the diagonal content of a note saying 'I'm sad' can be paraphrased as 'the person writing this was sad (at the time of writing)'. Diagonalization can be applied to both verbal and pictorial utterances (effectively turning (6) into something equivalent to our original pictorial content definition in (7)).

The following unified picture thus emerges: horizontal propositions model what is typically conveyed by an utterance in a direct face-to-face interaction where the context of creation is known to all involved, while diagonal propositions model what is typically conveyed by recorded, drawn, or written messages where the context of creation is not directly accessible to the interpreter.

In conclusion, written or spoken utterances and pictures all express very similar contents, modeled above in a familiar possible worlds and contexts framework. The way these propositions are expressed is different: verbal utterances have a hierarchical syntactic structure (generated with recursive rules and a lexicon of primitive elements) and a semantics that derives the propositional content from that structure via a system of composition rules and lexical meanings, while pictures have no syntactic structure and hence rely on a holistic semantics that derives propositional content by means of a single, general rule involving projection.

4 Pictorial storytelling

4.1 Discourse coherence and SDRT

A correct interpretation of a multi-sentence discourse typically includes more information than is contained in the interpretations of its individual sentences taken in isolation. Take the mini-discourse in (10).

(10) Macy was driving home late. A deer ran in front of her car. She swerved to avoid it and hit a tree. Her car was severely damaged.

We naturally infer that a deer ran in front of Macy's car *while* she was driving home late *and then* she swerved to avoid it *and then* she hit a tree and *as a result* her car was damaged. The sentences, or rather clauses, that make up this story strictly speaking only describe certain states and events occurring, but we as interpreters involuntarily combine those into a coherent discourse by inferring various causal, temporal and other relations between them (Hobbs 1979, Mann & Thompson 1988, Asher & Lascarides 2003). Such coherence inferences are generally defeasible and constrained by rationality, world-knowledge, a finite inventory of potential discourse relations (NARRATION, BACKGROUND, ELABORATION, EXPLANATION, etc.), and

linguistic cues (an overt connective like *and then* would signal NARRATION, *because* would signal EXPLANATION).

Segmented Discourse Representation Theory (SDRT) is a semantic framework offering an explicit formal model of these complex inferential processes. It treats each individual clause in a discourse as contributing a separate discourse unit, formulates a number of rules that constrain the inference of discourse relations between the linguistically encoded discourse units, and gives model-theoretic interpretations for the inferred relations. Below I will present a (highly simplified) SDRT-style model of discourse interpretation, breaking it down into four steps, and illustrating those by applying them to the short story above. These four steps are just a convenient presentation tool, not necessarily a cognitive reality. In the remainder of the paper I'll often refer back to these four steps in order to pinpoint where we need to make adjustments if we want the general SDRT theory of coherence-driven discourse interpretation to apply to pictorial narratives as well.

Step 1: Segmentation

The discourse is split into distinct discourse units, each (typically) describing a state or event and hence (typically) corresponding to a sentence or a clause – we'll ignore discourse management and other non-propositional discourse moves. In (10) we have the following five clausal discourse segments, describing five propositions, each describing the occurrence of an eventuality (i.e., a state or event).

(11) π_1 : Macy was driving home late.
π_2 : A deer ran in front of her car.
π_3 : She swerved to avoid it
π_4 : and hit a tree.
π_5 : Her car was severely damaged.

Step 2: Representing unit contents

We compute the contents of the individual discourse segments. The first, π_1, in our example expresses that there exists a person named Macy and an eventuality of her driving home late during some time interval in the past. We could represent these contents in some dynamic variant of first-order logic with eventualities, such as Discourse Representation Theory (DRT, Kamp 1981). Henceforth I'll use the notation $\langle\!\langle \pi_1 \rangle\!\rangle$ to denote the DRS representation of discourse segment labeled π_1. I'm skipping over the details of the compositional construction algorithm (mapping a syntactic parse of the input sentence to a so-called Preliminary DRS representing that sentence's context change potential), as well as the syntax and model-theoretic

semantics (notation: $[\![K]\!]$) of basic DRT. See Geurts et al. (2016) for a gentle introduction, and Kamp et al. (2003) for more details.

(12) $\langle\!\langle \pi_1 \rangle\!\rangle = \langle\!\langle$Macy was driving home late$\rangle\!\rangle =$

e_1 x_1
macy(x_1)
drive(e_1,x_1)
...

Due to the clause-level segmentation in Step 1, each of our five discourse units introduces and describes (among other things) a single main eventuality, which we'll represent with the correspondingly numbered discourse referents $e_1, \ldots, e_5$.

Step 3: Inferring discourse structure

We apply SDRT's system of defeasible inference rules to connect the discourse units to each other in a maximally coherent way. These rules might say that a sequence of a state description followed by an event description (like $\pi_1 - \pi_2$ here) allows the inference of a BACKGROUND connection, while a sequence of two events ($\pi_2 - \pi_3$) allows the inference of a NARRATION (under certain further conditions). With such rules (and some more global structural constraints) we can construct a number of potential discourse graphs for our story, including (13).

(13)
$$\begin{array}{l} \pi_1 \\ \Big\downarrow \text{BACKGROUND} \\ \pi_2 \xrightarrow{\text{NARRATION}} \pi_3 \xrightarrow{\text{NARRATION}} \pi_4 \xrightarrow{\text{RESULT}} \pi_5 \end{array}$$

We use horizontal arrows to visualize *coordinating* discourse relations (i.e., discourse relations like NARRATION and RESULT that in some intuitive sense move the story forward), and vertical edges to visualize *subordinating* relations (i.e., relations like EXPLANATION or BACKGROUND that explore subtopics of the 'dominant' node). The processes of inferring a discourse graph are intertwined with other inferences in the semantics/pragmatics interface, like presupposition and anaphora resolution, which are sensitive to discourse structural properties.

Step 4: Interpreting discourse graphs

From the various possible discourse graphs generated in Step 3 we need to choose the one that is maximally coherent, i.e., the one that has a maximal number of

connections, and that is overall semantically consistent and pragmatically most plausible. In order to assess the semantic and pragmatic optimality of a discourse graph we need a proper model-theoretic semantics, i.e., we have to determine the semantic content expressed by the entire connected discourse graph. We do this by combining the semantic interpretations of the elementary discourse units (Step 2) with semantic interpretation rules for the individual discourse relations, like (14). Notation: '$e_2 \prec e_3$' is a DRS condition that means that the eventuality described by π_2 immediately precedes that described by π_3 and $\oplus$ denotes the DRS equivalent of conjunction, viz. pairwise merge, i.e., creating a new DRS containing both conjuncts' discourse referents and conditions.

(14) $$\left\langle\!\left\langle\, \pi_2 \xrightarrow{\text{NARRATION}} \pi_3 \,\right\rangle\!\right\rangle = \langle\!\langle \pi_2 \rangle\!\rangle \oplus \langle\!\langle \pi_3 \rangle\!\rangle \oplus e_2 \prec e_3$$

The interpretation of an entire complex discourse graph goes through the construction of a big interpretable standard DRS that combines the information from all the individual units and from the ways they are connected.

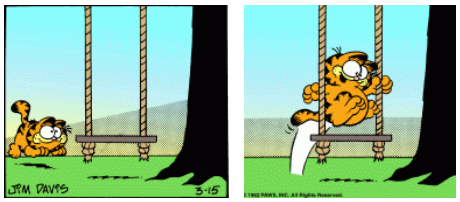
(15) $$\langle\!\langle \pi_i \rightarrow \pi_j \rightarrow \ldots \rangle\!\rangle = \langle\!\langle \pi_i \rightarrow \pi_j \rangle\!\rangle \oplus \langle\!\langle \pi_j \rightarrow \ldots \rangle\!\rangle$$

4.2 Pictorial discourse interpretation

SDRT models the interpretation of a multi-sentence discourse as a process guided by the inference of discourse relations between elementary discourse units ultimately leading to a maximally coherent interpretation. In our example story (10) the discourse segments corresponded to verbal clauses, each expressing a proposition about a main eventuality, introduced by the clause's main verb, and the inferred discourse relations included BACKGROUND, NARRATION and RESULT. But as we have seen, expressing propositions is not the prerogative of verbal signs; pictures express propositions as well. So it makes sense that a sequence of pictures could be interpreted as a coherent discourse in the same way as a sequence of sentences. I claim that this is exactly what happens when we read comics or graphic novels.

Consider the following two panels of a wordless Garfield strip.[2]

(16)



2 Jim Davis, *Garfield*, March 1982. PAWS Inc.

Intuitively, we interpret this as recounting a coherent sequence of two eventualities in temporal succession. We can tell a very similar (beginning of a) story with a sequence of English sentences:

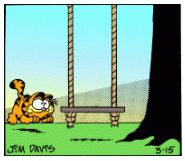
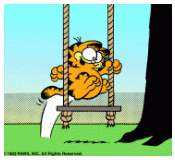
(17) Garfield was sitting in the grass, staring at a swing. Then he jumped onto the swing.

The hypothesis I explore here is that comics interpretation is driven by coherence maximization, just like verbal discourse interpretation, and moreover this interpretation process can be fruitfully analyzed via essentially the four-step algorithm outlined above. The difference with verbal storytelling is mostly confined to Step 2, the semantic interpretation of elementary discourse units. The construction and interpretation of the graph of discourse relations (Steps 3 and 4) should be essentially the same for both media, although we'll have to make a few modifications there as well.

To verify this hypothesis, let's see what happens when we try to apply our four-step process to the simple sequence in (16).

Step 1: Segmentation

Our starting assumption is that each panel is an elementary discourse unit. This is because, on the inherently holistic, projection-based semantics of section 3, a picture is the minimal unit that bears propositional content, i.e., that is *about* something, and hence can be judged true (accurate) or false (inaccurate). So:

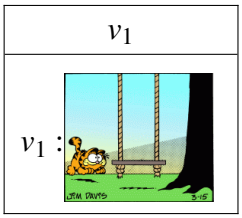
(18) π_1:  π_2: 

Step 2: Representing unit contents

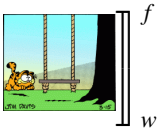
For the verbal case we settled on the dynamic, representational framework of DRT because we need the power of discourse referents to describe and explain discourse structural constraints on presupposition and anaphora resolution beyond the sentence (Asher & Lascarides 2003). As Abusch (2012) and Maier & Bimpikou (2019) argue, we need discourse referents in the analysis of pictorial narratives as well, in order to represent the inference that, say, the reddish cat-shapes in the panels of (16) denote the same cat. Let me briefly recap Maier & Bimpikou's (2019) DRT-based implementation of panel interpretation, called PicDRT.

Following Kamp (1981) we take DRSs to be structured mental representations that mediate between the discourse (verbal or pictorial) and the model-theoretic interpretation thereof (in terms of sets, functions, possible worlds, individuals, and truth values). Following Kosslyn (1980) I take it that mental representations themselves can be (partly) pictorial, which we might model by having actual pictures feature in DRS conditions.[3] Concretely, a so-called pictorial DRS condition couples a picture with a discourse referent representing the picture's viewpoint, leading to simple PicDRS boxes like (19).

(19) $\langle\!\langle \pi_1 \rangle\!\rangle =$



In the model-theoretic semantics of PicDRT such a condition can be interpreted relative to a model (which provides the projection function Π), a partial assignment function f, and a possible world w, by leveraging our projection-based semantics for pictures in a semantic interpretation rule like:

(20) $[\![v_1 :$  $]\!]^{f}_{w} = 1$ iff $\Pi(w, f(v_1)) =$ 

Adding (20) to the standard model-theoretic semantics of the DRS language, it follows that the PicDRS in (19) expresses that there is a viewpoint from which the world looks like like that picture (given the model's visual projection method).

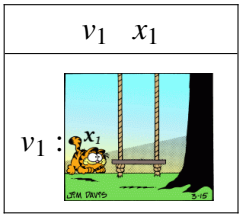
We assume further that in processing a picture the visual system identifies a number of salient 'regions of interest', i.e., parts of the picture that might be taken to represent the individuals that the picture (or visual narrative as a whole) is about. In other words, the visual counterparts of what dynamic semantics calls discourse referents. We model this part of the interpretation process formally by introducing fresh individual discourse referents as labels for the salient regions of interest in the picture under consideration. How exactly humans manage to isolate regions of interest in visual cognition is beyond the scope of this paper, and perhaps of semantics proper, but it's clear that it combines general cognitive

3 The inclusion of pictures in DRS conditions sets PicDRT apart from other (S)DRT accounts of pictorial narrative, like Bateman & Wildfeuer (2014) or Wildfeuer (2014), who replace pictures with symbolic descriptions of key features of their contents.

principles of Gestalt psychology, and artistic techniques for separating foreground and background, among other things.

In the case of the first Garfield panel, let's say the cat-shape is salient and will be labeled x_1, which gives the following basic DRS representation of that panel.

(21) $\langle\!\langle \pi_2 \rangle\!\rangle =$



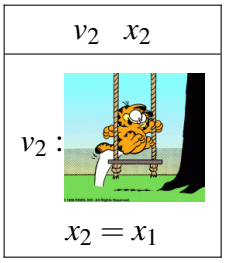
I refer to Maier & Bimpikou (2019) for a proper extension of the model-theoretic interpretation clause in (20) that links discourse referents with regions (i.e., $f(x_1)$, the individual picked out by discourse referent x_1, gets projected unto the correspondingly labeled cat-shape by Π from the viewpoint $f(v_1)$). With that in place, the rather minimal pictorial unit representation in (21) essentially captures the information that there is an individual and a viewpoint such that that individual (situated in its surrounding environment) looks like this, from that viewpoint. I take it that that is essentially all that the panel, in isolation, as an individual picture, conveys semantically.

Step 3: Inferring discourse relations (and other pragmatic enrichments)

Based on (discourse) context, common sense, world-knowledge, and discourse structural constraints, we can greatly enrich the minimal semantic contents of the individual units. In the presentation of verbal discourse interpretation above I focused on inferring discourse relations, but I also noted that this process is thoroughly intertwined with other pragmatic inference processes, like presupposition and anaphora resolution (Asher & Lascarides 2003). In the pictorial domain we have neither presupposition nor anaphora – there is no pictorial morphology that marks something as already given, old information. Hence, panel 2 in (16) could in principle be interpreted as introducing a new orange cat (Garfield's twin brother, perhaps). In context, the maximally coherent interpretation however is one where the two cat shapes in the two adjacent panels refer to the same cat. Maier & Bimpikou (2019) model this inference as the addition of an identity condition, $x_2 = x_1$, to the PicDRS. Note that, while Maier & Bimpikou (2019) use the standard, DRT model where each incoming sentence (or panel) is modeled as an update of a single, growing context DRS, we now use the SDRT model which represents the individual units separately while building a global discourse graph structure to represent their connections. Depending on the graph structure, different unit representation boxes may afford referential

dependencies (i.e., sharing discourse referents). Thus, the pragmatic inference that the two discourse referents introduced by the two orange regions represent the same individual can then be represented with an identity statement in the second box:

(22) $\langle\!\langle \pi_2 \rangle\!\rangle =$

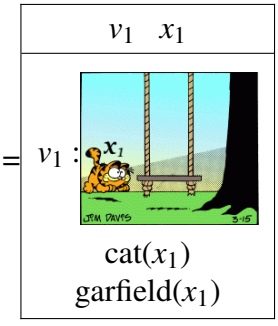


In contrast to the verbal domain, where we'd arrive at such an identity condition by resolving a pronoun or other definite noun phrase referring back to the previously established cat and swing (*he jumps onto the swing*), this condition is now the result of a process of 'free pragmatic enrichment', i.e., a pragmatic inference, based on context, world-knowledge, coherence, common sense, etc, that is not morphologically triggered, but may nonetheless affect the discourse's truth conditions (Recanati 2010).

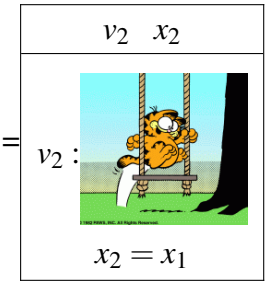
At this point we may also enrich the discourse representation under construction with further pragmatic inferences from context and world-knowledge, like the fact that x_1 is a cat, and that he is named 'Garfield', etc.. Finally, as in verbal discourse, Step 3 crucially involves the inference of discourse relations, which have been characterized as instances of free pragmatic enrichment as well (Pagin 2014). Intuitively, it seems very plausible to assume temporal progression, spatial proximity, and thematic continuity, so we'd want to add NARRATION. The end result of Step 3, after these various forms of pragmatic enrichment, will look roughly as in (23), where I first represent the discourse graph and then the semantics representations of the individual units:

(23) a. $\pi_1 \xrightarrow{\text{NARRATION}} \pi_2$

b. $\langle\!\langle \pi_1 \rangle\!\rangle =$



$\langle\!\langle \pi_2 \rangle\!\rangle =$



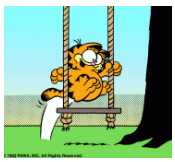
By our hypothesis, the inference and interpretation of discourse relations should be uniform across modalities, but when we try to apply the standard SDRT graph construction and interpretation mechanisms, we hit a snag: there are no discourse referents for states or events in (23). In the verbal domain, we assumed a Davidsonian analysis where elementary units are clauses with verbs that provide us with eventualities. This was crucial for two reasons: (i) the type of eventuality constrains the inference of discourse relations (Step 3), and (ii), many discourse relations are interpreted as relations between these eventualities (Step 4). But the Maier & Bimpikou (2019) analysis reconstructed above doesn't give us any eventualities. As it turns out, the question of whether pictures can express eventualities at all, and if so, what kinds, and how, is currently a subject of some debate, which I'll review and apply to our Garfield example in the next section.

5 Depicting events

5.1 Pictorial aspect and Abusch's Hypothesis

According to Abusch (2020), pictures are inherently stative; they literally depict a state of affairs, the way the world *is* at some point in space and time. We could implement Abusch's Hypothesis, as Altshuler & Schlöder 2021 call it, by simply introducing a fresh state discourse referent with each picture condition, i.e., an s_i alongside each v_i, referring to the state of affairs depicted in the picture.

For our first Garfield panel above, Abusch's Hypothesis makes sense. We can indeed plausibly interpret this picture as depicting a state of Garfield sitting on the grass while looking at a swing. However, for the second panel, (24a), a similarly stative paraphrase, (24b), feels less natural than an eventive one, (24c):

(24) a. 

b. ?Garfield {floats/is floating/is} in the air.

c. Garfield jumps in the air.

In addition to awkward stative paraphrasing, note that we readily infer narrative progression in the original sequence: Garfield's jump in panel 2 occurs *after* his staring at the swing in panel 1, and in close spatial and temporal proximity. As Schlöder & Altshuler (2023) point out, standard accounts of discourse semantics in the verbal domain associate narrative progression strictly with events, while

state descriptions are interpreted as temporally overlapping a previously mentioned eventuality (Kamp & Rohrer 1983, Altshuler 2016), as illustrated in (25).[4]

(25) Macy {left$_{event}$/was leaving$_{state}$} the hotel.
  a. She was$_{state}$ really drunk that night.
  b. She got$_{event}$ really drunk that night.

The stative continuation in (25a) cannot describe a situation were Macy left the hotel sober and then later that night got really drunk; she must have been drunk at the hotel. Only an eventive continuation like (25b) allows a reading with temporal progression, where the drinking occurs after she leaves. Transposing this observation to the Garfield comic, we get the following puzzle: if panel 2 in (16) is stative, as per Abusch's Hypothesis, where does the apparent temporal progression come from?

A number of recent papers explicitly address this pictorial aspect puzzle: (i) Abusch (2020) denies that temporal progression requires an event and assumes that temporal sequencing is hardwired into pictorial sequencing[5]; (ii) Altshuler & Schlöder (2021) argue that discourse relations work differently across different media, with Schlöder & Altshuler (2023) subsequently devising a multimodal discourse interpretation system where pictures differ from verbal sentences in not introducing discourse referents for eventualities at all; and (iii) Maier (2019) and Schlöder & Altshuler (2023) briefly mention, but don't develop or endorse, a purely pragmatic view where eventuality discourse referents are introduced at the level of pragmatic enrichment.

Below I want to explore an alternative, based on Maier (2019):pictures semantically introduce eventuality discourse referents, but further aspectual specification is left to either pragmatic enrichment or special visual aspect markers.

5.2 Introducing events

Maier (2019) argues that pictures are not always stative, pointing to, among other things, comics panels with speech bubbles as unambiguous depictions of (speech) events. Instead, every individual identified as such in the picture (as a salient region) is assumed to be a participant in some eventuality. The idea behind this is that in pictures we can see eventualities through their physical participants (agents, patients,

4 The state–event sequence is a bit awkward, but to the extent that it's interpretable it clearly allows a narrative progression reading. Also observe that we can felicitously add markers forcing temporal progression without making the discourse awkward: Macy was leaving the hotel. And then later that night she got really drunk.

5 Maier & Bimpikou (2019) likewise assume default temporal progression for pictorial sequences, without relying on event representations, while Fernando (2020) goes one step further and derives temporal progression from pictorial sequencing as a form of iconicity.

themes, experiencers, etc). Moreover, every individual we see can be safely assumed to do, be, experience, receive, or undergo something. Formally, along with each individual discourse referent x_i we introduce an eventuality discourse referent e_i and a condition to the effect that x_i participates in e_i (where 'participates' is a highly underspecified relation encompassing all thematic roles). For the first two panels that gives us (at the minimal semantic level, Step 2, i.e., before pragmatic enrichment):

(26) $\langle\!\langle \pi_1 \rangle\!\rangle =$ $\begin{array}{|c|}\hline v_1 \quad x_1 \quad e_1 \\ \hline v_1 : \\ \text{partcpt}(x_1, e_1) \\ \hline \end{array}$ $\langle\!\langle \pi_2 \rangle\!\rangle =$ $\begin{array}{|c|}\hline v_2 \quad x_2 \quad e_2 \\ \hline v_2 : \\ \text{partcpt}(x_2, e_2) \\ \hline \end{array}$

This already brings the form of our elementary discourse unit representations more in line with what we need to interpret discourse relations as relations between eventualities (Step 4). Moreover, we're no longer assuming that each picture is stative so it's indeed possible to infer a NARRATION or RESULT relation (Step 3) and thus capture the intuitive forward movement of the narrative.

Since the semantics leaves it underspecified whether we're dealing with a state or an event, the space of possible discourse relations between any two panels is rather unconstrained. We can of course appeal to pragmatics to infer that the eventuality that Garfield participates in is an event, more specifically a jumping event, and that Garfield's mode of participation is agentive. Such aspectual inferences are then entirely on a par with the other forms of pragmatic enrichment in Step 3, such as that the salient individual is a cat named Garfield, and that he is the same cat as that depicted in the first panel.

When there are multiple salient individuals depicted in a single panel we might introduce multiple eventuality discourse referents. Some might actually be equated, as multiple individuals could be participating in a single eventuality, but some might still refer to distinct events. At that point we might have to pragmatically single out one as the main eventuality.[6] In some cases, especially some genres or works without clear panel borders (like the famous 11th century Bayeux tapestry), it might be necessary to split a single large picture into two or more elementary discourse units, but I will leave this for a future occasion and focus here on the simple Garfield comic.

We've noted above that in Step 3 the inference of discourse relations is thoroughly intertwined with other forms of pragmatic enrichment. Inferred identities between discourse referents and aspectual specification, for instance, may both influence the

6 Alternatively, or additionally, we could combine multiple eventualities to infer a complex 'supereventuality' and take that as the main eventuality depicted.

choice of discourse relation, but at the same time discourse structural considerations may inform inferences of identity and eventuality specification. I will present just the end result of the intertwined semantic and pragmatic processing under Step 3 for our Garfield panels:

(27) a. $\pi_1 \xrightarrow{\text{NARRATION}} \pi_2$

b. $\langle\!\langle \pi_1 \rangle\!\rangle =$

v_1 x_1 e_1
v_1: 
partcpt(x_1, e_1)
cat(x_1)
garfield(x_1)
state(e_1)
crouch(e_1)

$\langle\!\langle \pi_2 \rangle\!\rangle =$

v_2 x_2 e_2
v_2 : 
agent(x_2, e_2)
$x_2 = x_1$
event(e_2)
jump(e_2)

If we now apply Step 4, the model-theoretic discourse semantics, to (27) we get the following interpretation: first there's a state of Garfield being crouched down, which looks like panel 1 from some viewpoint v_1, and then there's an event of him jumping, which looks like panel 2 from some viewpoint. In addition we could always add further pragmatic enrichments, like an additional discourse referent y_1 for the swing, and another state referent s_1 of Garfield staring at that swing. Different from Abusch (2020) and Maier & Steinbach (2022), the temporal order between the two eventualities is encoded in the model-theoretic interpretation rule for NARRATION(see Section 4.1).

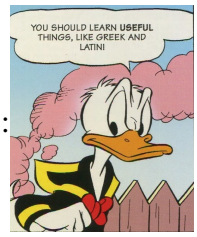
Summing up the account so far: sentences and pictures both express propositions and when put in sequence, both modalities (and mixtures) can be used to tell coherent stories. SDRT is a suitable framework for modelling the interpretation processes involved in the interpretation of storytelling in both the verbal and the pictorial domain. The overall architecture of SDRT applies to both but we have encountered some differences: (i) pictures express propositions in a different way – via projection rather than composition; (ii) pictures can introduce discourse referents for individuals and eventualities, but these are highly underspecified, i.e., pictorial storytelling happens without pictorial analogues of such ubiquitous functional linguistic markers as pronominal features, tense, aspect, or modality; and (iii), as a result of that semantic underspecification, pictorial narrative interpretation relies heavily on various forms of pragmatic enrichment, from the identification of individuals across panels, and the attribution of (non-visual) properties of individuals and eventualities, to the inference

of discourse relations. In the following we'll examine some candidates for pictorial event markers in the semantics that can take some of the load off pragmatics.

5.3 Speech bubbles

Photos, seventeenth century portrait paintings, and many drawings express their contents via geometric projection. We've seen how coherent stories can be communicated by putting such more or less realistic pictures in sequence. A lot of information relevant to the extraction of a coherent plot is then not directly encoded but left to pragmatics. However, comics artists today are actually relying on a rich inventory of conventionalized markings to encode some of this information directly in the panel. In the remainder of this section I take a brief look at some of these.

Consider a simple example of a speech bubble, common in many comics genres (Cohn 2013a).[7]

(28) π_1: 

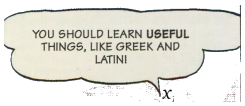
The so-called tail of the speech balloon usually points towards a salient individual and indicates that that person is the agent of a speech act. The words inside the balloon display the words spoken by the agent.

Crucially, the speech balloon is not (usually) interpreted projectively. It's a convention that instructs the reader to imagine a speech act, not a floating white balloon. We thus add another step to the interpretation process, where the panel is decomposed into a basic pictorial element, interpreted projectively, and the speech balloon, interpreted symbolically. I'll assume that in the interpretation of a panel, the picture–symbol decomposition occurs at the pre-semantic phase, where we also identify the visually prominent regions and introduce corresponding discourse referents – in Step 2. For simplicity, I'll assume that at this stage the balloon's tail can already be identified as pointing to one such salient region, and hence can be identified with a specific discourse referent.[8]

7 Panel from *Donald Duck on Treasure Island*, Disney.

8 If the tail cannot be deictically identified with a salient region in the current panel we introduce a new discourse referent and leave it to the later pragmatic enrichment to figure out who or what this speaker is and equate the corresponding discourse referents.

(29) a. 

b. 

In sum, we view a panel like (28) as a picture–symbol hybrid which we have to decompose so we can apply distinct interpretation strategies to the two semiotic components.

First, the pictorial component, (29a). After decomposition, the picture will have a 'hole' where the balloon sat, i.e., a region where no information is given about what the world looks like. The possible worlds approach to pictorial meaning is exceptionally well-suited for dealing with this kind of underspecification. We can implement the interpretation of holey pictures by assuming a holey picture plane where projection lines pass through without triggering any markings.

The speech balloon, (29b), meanwhile, can be interpreted through an arguably symbolic interpretation rule, a close cousin of the rule for interpreting written language quotation. To make this precise I follow a standard event-based semantics of quotation (Maier 2017). The balloon introduces a new event discourse referent for a speech act, and the contents of the balloon approximate (in the written modality) the linguistic surface form of the words spoken. Formally, the condition 'form(e,...)' demands that event e (i.e., the speech act) has a phonetic form approximated orthographically by the quoted string of letters. The agent of the speech act is also specified, viz. as the individual discourse referent associated with the tail.

(30) $\langle\!\langle$ 

$\rangle\!\rangle$ =

e
say(e)
form(e, 'You should learn useful things, like Greek and Latin!')
agent(e,x_1)

Now we can combine the two component contributions to get the full meaning of the panel:

(31) $\langle\!\langle \pi_1 \rangle\!\rangle =$  $\oplus$  =

v_1 x_1 e_1 e
v_1: 
partcpt(e_1, x_1)
say(e)
form(e, 'You should learn useful things, like Greek and Latin!')
agent(e, x_1)

Speech balloons are treated as symbolic enrichments of a picture. They enrich the panel's semantic content with fairly specific information that is quite hard to convey pictorially and thus greatly expand the depth and type of stories that can be told.[9]

The proposed decompositional analysis could easily be extended to cover speech and thought balloons, containing text or further pictures (thus allowing for recursive embedding of further balloons etc.). I leave a proper investigation for another occasion.

5.4 Symbolic enrichments

I want to end this chapter by exploring a generalization of the above to a general decompositional strategy for dealing with picture–symbol hybrids, including Cohnian affixes as discussed in Section 3.1, exemplified (3), repeated in (32).

(32) 

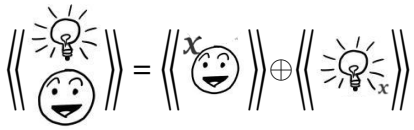
Drawing a light bulb symbol over a character's head to indicate inspiration is part of a well-known set of conventions that comics artists in some contemporary genres draw on to convey their characters' mental states. The light bulb convention is not quite as arbitrary as the lexical rule that connects the English word 'inspiration' to its meaning. It has its roots in an easily reconstructed metaphor, viz., inspiration is

9 Maier (2019) further decomposes the text balloon itself into two separate semiotic elements: the text is treated as a piece of quoted text in the written modality and the balloon can be viewed as part indexical (the tail points towards an individual region) and part symbol (for instance, we have the more or less arbitrary convention that white ellipses indicate speech, while cloud-like bubbles indicate private thoughts, and spiky shapes might indicate yelling).

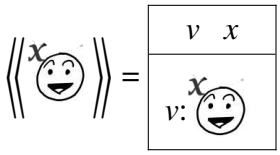
like shining a light in the darkness (to uncover the unknown, presumably). But like a speech balloon it's clearly not iconic, let alone pictorial, either, as we don't usually take the picture to entail the existence of (something that resembles) an incandescent light bulb floating right above the person's head. The light bulb is an example of a 'motivated symbol' (Greenberg 2022), 'translucent sign' (Davidson 2015), or 'descriptive icon' (Davidson 2023), i.e., a sign that gets its meaning based on a non-natural, learned convention, stored in the lexicon, but one that is not arbitrary and that, once learned, can be easily explained and remembered.

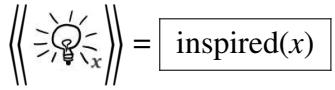
The lexicons of sign languages are a rich source of examples of translucent signs. You can't predict a priori what the ASL sign for cat is, and indeed it varies across different sign languages, but once taught – an F handshape moving out from the side of your mouth while pinching thumb and index finger together – it 'makes sense' – you're drawing a cat's whiskers. Following both Davidson and Greenberg, I take translucent signs to be symbols, semantically speaking, despite their straightforward etymological explanations.

Now, if we assume the interpreter recognizes the light bulb as a conventional symbol, then it will gets separated and processed symbolically, much like a speech balloon. Specifically, we decompose the picture into (i) a depiction of a face, labeled with a discourse referent x, and with a hole above the head where the symbol was, and (ii) a lexical element, tagged with that same x.

(33) 

The first component gets interpreted via projection, the second via a stored lexical semantic rule for lightbulb symbols.

(34) a. 

b. 

5.5 Pictorial aspect revisited

Now, finally, we return to the issue of visual aspect marking. Our current analysis of speech balloons refutes (an uncharitably strong reading of) Abusch's Hypothesis in that any panel with a speech balloon describes a speech event and hence such pictures are not semantically stative. The defender of Abusch's Hypothesis of course

might respond and call such panels multimodal hybrids rather than pictures, perhaps even if the speech balloon contains an image rather than words, as in (35).[10]

(35) a.  b.  c. 

Now take another look at the entirely wordless, balloonless Garfield strip in (16), especially panel 2, repeated in (36) (annotated with discourse segment and region labels).

(36) π_2: 

In Section 5.1 I argued on the grounds of pragmatic reasoning involving discourse structure, coherence, and world-knowledge that this second panel depicts an event of Garfield jumping, rather than a state of Garfield being in the air. A closer look reveals some morphological support for this interpretation, viz. the white arc that, intuitively, serves to indicate the path of Garfield's movement and hence arguably entails a change of state – an event. Thus, (36) is also a counterexample to Abusch's Hypothesis, as, interestingly, Abusch herself would agree:[11]

> I expect that if a semantics for motion lines were added to the semantic basis, some panels would prove to have non-cumulative propositional content. This could revive the possibility of an aspectually sensitive construction rule. (Abusch 2014)

Let me end the chapter with some notes on what such a construction rule would (not) look like in the current framework.

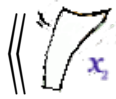
At first sight, motion lines share some characteristics with symbolic enrichments like the lightbulb affix. Like other symbolic affixes they are quite common and readily interpretable by contemporary readers, and come in many specific conventionalized forms that have various more or less specific interpretations across comics genres (Hacımusaoğlu & Cohn 2022). Crucially, an image like (36) does not depict Garfield attached to a white arc sprouting from the grass, with two little black curved lines

10 Panels from Doxiadis et al *Logicomix* (Bloomsbury, 2009), Nathan Pyle (Facebook, February 27, 2020), Jim Davis *Garfield* (PAWS Inc., September 9, 1987).

11 Abusch (p.c.) suggests that Schlöder and Altshuler misattribute the eponymous hypothesis to her.

floating under his tail. Motion lines are not actually there in Garfield's world, they are more like morphemes that conventionally indicate types of movements.

In the current framework we could try to go the decomposition route again and treat motion lines as symbolic enrichments. The white arc in (36) would then get separated from the surrounding image and be subject to a learned lexical rule like (37), which encodes that the anchor (in this case represented by the discourse referent x_2) is moving.

(37) $\langle\!\langle$  $\rangle\!\rangle$ =

e'
move(e') agent(e',x_2)

Full picture–symbol decomposition along these lines is ultimately unsatisfactory. The motion line doesn't just encode *that* there is movement, it also *shows* the path and type of movement. Although we can't interpret the picture holistically (depicting Garfield on a white arc in the grass), we also can't just treat the arc as a separable symbol, because the shape of the arc depicts the path of the movement. In fact, we can't even treat it as a separable icon, i.e., depicting the shape of the movement iconically, because we actually get information about the starting point, and the path of the movement in space, i.e., relative to the grass, the swing, and the tree. The challenge of motion line interpretation is one of finding a balance between separating the line from the projectively interpreted host picture, while nonetheless anchoring the shape of the motion line to the depicted world.[12]

6 Conclusion

Sentences and pictures both express propositions. They are used to represent the (actual or fictional) world as being a certain way. The propositions thus expressed can be fruitfully modeled uniformly as sets of possible worlds, but the way these propositions are expressed differs: sentence interpretation relies on a recursive grammar and compositional semantics, while picture interpretation relies on geometric projection.

When propositional units of either modality are put in sequence they can be used to tell coherent stories, as shown in oral storytelling, novels, comics, and illustrated books. I've explored to what extent SDRT is a suitable framework for modelling the interpretation processes involved in the interpretation of storytelling in both the verbal and the pictorial domain. At first sight, the overall architecture of SDRT seems

12 An alternative route would be to treat motion lines as truly depictive after all, viz. as a kind of stylization of the motion blur that results from a long exposure in photography. Exploring and comparing these two suggestions will have to await future research.

to apply to both modalities but we have encountered some differences: (i) pictures express propositions in a different way – via projection rather than composition – which means we have to extend the semantics of basic DRT with projections; (ii) pictures can introduce discourse referents for individuals and eventualities, but these are highly underspecified, i.e. pictorial storytelling happens without pictorial analogues of such ubiquitous functional linguistic markers as pronominal features, case, tense, mood, etc.; and (iii), as a result of that semantic underspecification, pictorial narrative interpretation relies quite heavily on various forms of pragmatic enrichment, from the identification of individuals across panels and the attribution of (non-visual) properties of individuals and eventualities, to the inference of discourse relations.

A closer look at comics shows the emergence of some visual morphology that helps take some of the load off pragmatic enrichment and thus facilitates more complex storytelling. I considered here three cases in point: (i) speech balloons, treated as a visual analogues of quotation; (ii) visual affixes like the idea light bulb, treated as symbolic enrichment; and (iii) motion lines. The first two, or perhaps all three, mix pictorial and symbolic elements, which we should separate before we can semantically interpret them. The first and the last I use also to argue against Abusch's Hypothesis, i.e., the claim that pictures are by definition stative (as opposed to eventive).

The handful of examples discussed in the second half of the paper are all from pop culture comics, but many of the conclusions and observations generalize to other forms of (partly) pictorial communication. Take emojis, ubiquitous in text-based communication: so-called activity emojis, like 🏃, are arguably interpreted as denoting events, and many face emojis, like 😍, arguably contain both symbolic and iconic elements (Maier 2023, Grosz et al. 2023). Another case in point is film, a form of pictorial discourse that consists of a number of pictorial units, known as shots, put in a deliberate sequence. As in comics we can then think of the interpreter as inferring the relevant coherence relations between the events depicted in adjacent shots to extract a coherent story, leading us towards a view on film semantics continuous with comics and verbal discourse interpretation (Cumming et al. 2017, Wildfeuer 2014).

References

Abusch, Dorit. 2012. Applying Discourse Semantics and Pragmatics to Co-reference in Picture Sequences. *Proceedings of Sinn und Bedeutung* 17. http://ecommons.cornell.edu/handle/1813/30598.

Abusch, Dorit. 2014. Temporal succession and aspectual type in visual narrative. In Luka Crnič & Uli Sauerland (eds.), *The Art and Craft of Semantics: A Festschrift*

for Irene Heim, 9–29. MITWPL.
Abusch, Dorit. 2020. Possible-Worlds Semantics for Pictures. In *The Blackwell Companion to Semantics*, Wiley. https://doi.org/10.1002/9781118788516.sem003.
Altshuler, Daniel. 2016. *Events, States and Times: An essay on narrative discourse in English*. Walter de Gruyter.
Altshuler, Daniel & Julian Schlöder. 2021. If pictures are stative, what does this mean for discourse interpretation? *Proceedings of Sinn und Bedeutung* 25. 19–36. https://doi.org/10.18148/sub/2021.v25i0.922. https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/922.
Asher, Nicholas & Alex Lascarides. 2003. *Logics of Conversation*. Cambridge: Cambridge University Press.
Bateman, John & Janina Wildfeuer. 2014. A Multimodal Discourse Theory of Visual Narrative. *Journal of Pragmatics* 74. 180–208. https://doi.org/10.1016/j.pragma.2014.10.001.
Blutner, Reinhard, Helen De Hoop & Petra Hendriks. 2006. *Optimal communication*. Stanford: CSLI Press.
Chomsky, Noam. 1957. *Syntactic structures*. The Hague: Mouton and Co.
Cohn, Neil. 2013a. Beyond speech balloons and thought bubbles: The integration of text and image. *Semiotica* 2013(197). 35–63. https://doi.org/10.1515/sem-2013-0079.
Cohn, Neil. 2013b. *The Visual Language of Comics: Introduction to the Structure and Cognition of Sequential Images.* A&C Black.
Cumming, Samuel, Gabriel Greenberg & Rory Kelly. 2017. Conventions of Viewpoint Coherence in Film. *Philosophers' Imprint* 17(1). 1–29. http://hdl.handle.net/2027/spo.3521354.0017.001.
Davidson, Kathryn. 2015. Quotation, demonstration, and iconicity. *Linguistics and Philosophy* 38(6). 477–520. https://doi.org/10.1007/s10988-015-9180-1.
Davidson, Kathryn. 2023. Compositionality and Iconicity Ms. Harvard. https://projects.iq.harvard.edu/sites/projects.iq.harvard.edu/files/meaningandmodality/files/compositionality_and_iconicity.pdf.
Dingemanse, Mark. 2012. Advances in the Cross-Linguistic Study of Ideophones. *Language and Linguistics Compass* 6(10). 654–672. https://doi.org/10.1002/lnc3.361.
Fernando, Tim. 2020. Pictorial narratives and temporal refinement. *Semantics and Linguistic Theory (SALT)* 29.
Frege, Gottlob. 1892. Ueber Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik* 100(1). 25–50.
Geurts, Bart, David Beaver & Emar Maier. 2016. Discourse Representation Theory. In Edward Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University spring 2016 edn. https://plato.stanford.edu/

archives/spr2016/entries/discourse-representation-theory/.
Greenberg, Gabriel. 2013. Beyond Resemblance. *Philosophical Review* 122(2). 215–287. https://doi.org/10.1215/00318108-1963716.
Greenberg, Gabriel. 2022. The Iconic-Symbolic Spectrum. *Philosophical Review* .
Grice, H. P. 1957. Meaning. *The Philosophical Review* 66(3). 377. https://doi.org/10.2307/2182440.
Grosz, Patrick Georg, Gabriel Greenberg, Christian De Leon & Elsi Kaiser. 2023. A semantics of face emoji in discourse. *Linguistics and Philosophy* https://doi.org/10.1007/s10988-022-09369-8.
Hacımusaoğlu, Irmak & Neil Cohn. 2022. Linguistic typology of motion events in visual narratives. *Cognitive Semiotics* 15(2). 197–222. https://doi.org/10.1515/cogsem-2022-2013.
Hagen, Margaret. 1986. *Varieties of Realism: Geometries of Representational Art*. Cambridge: Cambridge University Press.
Heim, Irene & Angelika Kratzer. 1998. *Semantics in Generative Grammar*. Oxford: Blackwell.
Hendriks, Petra & Helen De Hoop. 2001. Optimality theoretic semantics. *Linguistics and philosophy* 24(1). 1–32.
Hobbs, Jerry. 1979. Coherence and Coreference. *Cognitive Science* 3(1). 67–90. https://doi.org/10.1207/s15516709cog0301_4.
Kamp, Hans. 1981. A theory of truth and semantic representation. In Jeroen Groenendijk, Theo Janssen & Martin Stokhof (eds.), *Formal Methods in the Study of Language*, 277–322. Amsterdam: Mathematical Centre Tracts.
Kamp, Hans. 2015. Using Proper Names as Intermediaries Between Labelled Entity Representations. *Erkenntnis* 80(2). 263–312. https://doi.org/10.1007/s10670-014-9701-2.
Kamp, Hans, Josef van Genabith & Uwe Reyle. 2003. Discourse Representation Theory. In Dov Gabbay & Franz Guenthner (eds.), *Handbook of Philosophical Logic*, vol. 10, 125–394. Heidelberg: Springer.
Kamp, Hans & Christian Rohrer. 1983. Tense in Texts. In *Meaning, Use, and Interpretation of Language*, 250–269. De Gruyter. http://doi.org/10.1515/9783110852820.250.
Kaplan, David. 1989. Demonstratives. In Joseph Almog, John Perry & Howard Wettstein (eds.), *Themes from Kaplan*, 481–614. New York: Oxford University Press.
Kosslyn, Stephen Michael. 1980. *Image and Mind*. Harvard University Press.
Kulvicki, John. 2013. *Images*. Routledge.
Lande, Kevin J. 2020. Mental Structures. *Noûs* (3). 649–677. https://doi.org/10.1111/nous.12324.

Lande, Kevin J. 2024. Pictorial Syntax. *Mind and Language* forthcoming. https://doi.org/10.1111/mila.12497.

Maier, Emar. 2017. The pragmatics of attraction: Explaining unquotation in direct and free indirect discourse. In Paul Saka & Michael Johnson (eds.), *The Semantics and Pragmatics of Quotation*, Berlin: Springer. http://ling.auf.net/lingbuzz/002966.

Maier, Emar. 2019. Picturing words: the semantics of speech balloons. *Proceedings of the Amsterdam Colloquium* 22. 584–592. https://philpapers.org/rec/MAIPWT.

Maier, Emar. 2023. Emojis as pictures. *Ergo* 10(1). https://doi.org/https://doi.org/10.3998/ergo.4641.

Maier, Emar & Sofia Bimpikou. 2019. Shifting perspectives in pictorial narratives. *Sinn und Bedeutung* 23(2). 91–106. https://doi.org/10.18148/sub/2019.v23i2.600.

Maier, Emar & Markus Steinbach. 2022. Perspective Shift Across Modalities. *Annual Review of Linguistics* 8(1). 59–76. https://doi.org/10.1146/annurev-linguistics-031120-021042.

Mann, William & Sandra Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text* 8(3). 243–281.

Montague, Richard. 1973. The Proper Treatment of Quantification in Ordinary English. In *Approaches to Natural Language*, vol. 49, 221–242. Dordrecht: Reidel.

Pagin, Peter. 2014. Pragmatic Enrichment as Coherence Raising. *Philosophical Studies* 168(1). 59–100. https://doi.org/10.1007/s11098-013-0221-8.

Peirce, C. S. 1868. On a New List of Categories. *Proceedings of the American Academy of Arts and Sciences* 7. 287–298.

Recanati, Francois. 2010. Pragmatic Enrichment. In Delia Fara & Gillian Russell (eds.), *Routledge Companion to Philosophy of Language*, 67–78. Routledge.

Saussure, Ferdinand de. 1916. *Cours de linguistique générale*. Lausanne: Payot.

Schlöder, Julian J. & Daniel Altshuler. 2023. Super Pragmatics of (linguistic-)pictorial discourse. *Linguistics and Philosophy* https://doi.org/10.1007/s10988-022-09374-x.

Shin, Sun-Joo, Oliver Lemon & John Mumma. 2018. Diagrams. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University winter 2018 edn. https://plato.stanford.edu/archives/win2018/entries/diagrams/.

Stalnaker, Robert. 1978. Assertion. In Peter Cole (ed.), *Syntax and Semantics 9: Pragmatics*, 315–332. New York: Academic Press.

Wildfeuer, Janina. 2014. *Film Discourse Interpretation: Towards a New Paradigm for Multimodal Film Analysis*. Routledge.

Willats, John. 1997. *Art and representation: new principles in the analysis of pictures*. Princeton, N.J.: Princeton University Press.

Wilson, Brent & Marjorie Wilson. 1977. An Iconoclastic View of the Imagery Sources in the Drawings of Young People. *Art Education* 30(1). 4–11. https://doi.org/10.1080/00043125.1977.11649876.

Zimmermann, Thomas Ede. 1991. Kontextabhängigkeit. In Arnim von Stechow & Dieter Wunderlich (eds.), *Semantik: Ein internationales Handbuch der zeitgenössischen Forschung*, 156–229. Berlin/New York: Walter de Gruyter.