Picturing words: the semantics of speech balloons[*]

Emar Maier

Theoretical Philosophy/MILLC, University of Groningen
e.maier@rug.nl

Semantics traditionally focuses on linguistic meaning. In recent years, the Super Linguistics movement has tried to broaden the scope of inquiry in various directions, including an extension of semantics to talk about the meaning of pictures. There are close similarities between the interpretation of language and of pictures. Most fundamentally, pictures, like utterances, can be either true or false of a given state of affairs, and hence both express propositions (Zimmermann, 2016; Greenberg, 2013; Abusch, 2015). Moreover, sequences of pictures, like sequences of utterances, can be used to tell stories. Wordless picture books, comics, and film are cases in point. In this paper I pick up the project of providing a dynamic semantic account of pictorial story-telling, started by Abusch (2012) and continued by Abusch & Rooth (2017); Maier & Bimpikou (2019); Fernando (2020).

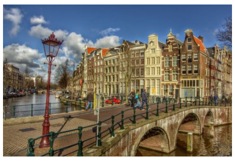
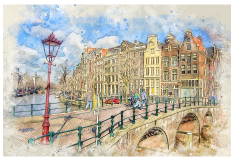
More specifically, I propose here a semantics of speech and thought bubbles by adding event reference and event modification to PicDRT. To get there I first review the projection-based semantics for pictures (section 1), noting the fundamental distinction between symbolic and iconic meaning that makes speech bubbles especially interesting (section 2). I then review the dynamic PicDRT framework for pictorial narratives (section 3), add events (section 4), and propose an account of speech bubbles as quotational event modification (section 5). I end with a brief look at other conventional event modifiers in comics such as motion lines (section 6).

1 Picture semantics

As a first pass at bringing our formal semantics toolkit to the study of pictures, we might say a picture expresses the proposition consisting of the set of all those worlds that 'look like' the picture. However, just as sentences with indexical elements express propositions only relative to a context of utterance (Kaplan, 1989), so pictures are usually assumed to express propositions relative to a viewpoint (formally, a unit vector located somewhere in space time, specifying from where, when, and in what direction we're observing/picturing the world). With this notion of a viewpoint we can introduce the notion of a projection function, which will then replace the vague (and problematic, Goodman 1976; Greenberg 2013) notion of resemblance in our first rough characterization. In short, a projection is a recipe for turning a 3D scene (part of a world seen from a viewpoint) into a 2D representation. For example, using linear perspective digital photography projection, as implemented in my phone, we can project the actual world w, from a viewpoint v facing the Keizersgracht from the bridge onto a digital photo, (1a). With a different projection algorithm, π_2, treating projection lines coming from edges and surfaces somewhat differently, we can map the same world, from more or less the same viewpoint, onto a 2D drawing:[1]

[*] Many thanks to Sofia Bimpikou and the anonymous AC reviewers, and to the workshop organizers. This research is supported by NWO Vidi Grant 276-80-004 (FICTION).

[1] Photo from https://pixabay.com/photos/amsterdam-keizersgracht-netherlands-686460/, drawing from https://pixabay.com/illustrations/amsterdam-keizersgracht-netherlands-3609378/.

(1) a. $\pi_1(w, v) =$  b. $\pi_2(w, v') =$ 

We can now more precisely define the proposition expressed by a picture as the inverse of the relevant projection, say π_2 for the interpretation of the drawing (Abusch, 2015).

(2) $[\![$ $]\!]^{v'} = \left\{ w \;\middle|\; \pi_2(w, v') = \right.$ $\}$

In other words, the meaning of a picture, relative to a fixed viewpoint v and projection function π provided by the model, is the set of worlds w that could be projected onto that picture from viewpoint v. However, when we interpret a picture we often don't have any prior, independent access to the spatio-temporal viewpoint the artist used to create her projection (this is especially, but not exclusively, true for fictional depictions). Instead, we may infer certain properties of the viewpoint from the picture itself, just like in interpretation we try to infer properties of the world depicted. To model this we use a more abstract, two-dimensional, or centered, notion of content, akin to a Kaplanian character, or its diagonal. Generally, for any picture α, interpreted in a model providing a projection function π alongside a set of worlds and a set of viewpoints (Rooth & Abusch, 2017):

(3) $[\![\alpha]\!] = \{\langle w, v\rangle \mid \pi(w, v) = \alpha\}$

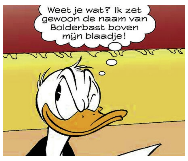
2 Iconic vs. symbolic meaning

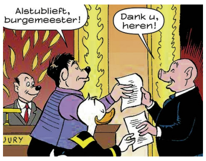
The key differences between pictorial and linguistic meaning are in *how* they express propositions. Linguistic meaning is considered mostly *symbolic* (i.e., depending on an arbitrary, purely conventional lexicon), and *compositional* (i.e., depending on a grammar specifying how meanings of complex expressions are built up out of constituent meanings). Pictorial meaning by contrast is mostly *iconic*, i.e. representing by virtue of some more or less natural, structure preserving transformation relation (like our projection above), and *holistic*, i.e. independent of grammatical constituent structure.

However, on closer inspection, many aspects of language turn out to be more or less iconic (onomatopoeia (Henderson, 2016) , co-speech gestures (Schlenker, 2018) , sign language classifiers (Davidson, 2015), etc.). Quotation and quotatives, especially in spoken language, are often analyzed as iconic as well: in quoting, a reporter 'demonstrates' a previous speech act by producing something 'similar' (Clark & Gerrig, 1990; Recanati, 2001; Davidson, 2015).

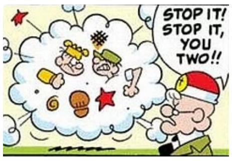
On the other side, the degree to which certain drawing or even photography and film styles are purely iconic has also been debated. Linear perspective drawing, for instance, is arguably not so 'natural' that early medievals were able to apply it. Moreover, it's not clear that the visual system itself follows linear perspective projection in the interpretation of our surroundings (Giardino & Greenberg, 2015; Greenberg, 2013).

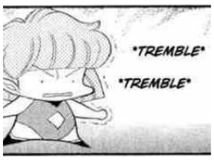
A rather different instance of symbolic/compositional meaning intruding in the pictorial domain comes (again, surprisingly) from quotation. If in a drawing we want to depict someone saying something, we often add writing. A common convention, especially in comics and cartoons, is to use the speech balloon symbol to show who said what (4b). Similarly, a thought bubble shows what a character is thinking (4a).[2]

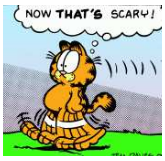
(4) a. 
Weet je wat? Ik zet gewoon de naam van Bolderbast boven mijn blaadje!

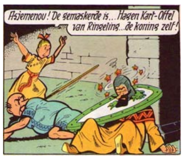
b. 

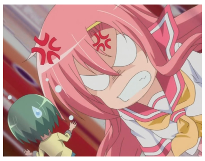
Consider also fight clouds (5a), sound effect descriptions (5b), explosion stars, motion lines/reduplication/blurring (5c), and what Cohn (2013) calls 'affixes', like the idea-lightbulb, angry-thundercloud, stars circling over a passed out character's head (5d), or the bulging-vein anger symbol (5e). All of these symbols serve as conventional indicators of certain types of events or (mental) states, significantly enriching the purely projective, pictorial meaning of the panels in which they occur.

(5) a. 

b. 

c. 

d. 
Asjemenou! De gemaskerde is... Hagen Kart-Offel van Ringeling...de koning zelf!

e. 

The goal of this paper is to give a compositional semantics of speech and thought balloons, and integrate that into a projection-based dynamic semantic account for pictorial narratives. In the final section I share some preliminary thoughts about some of the other symbolic enrichments in (5).

3 PicDRT

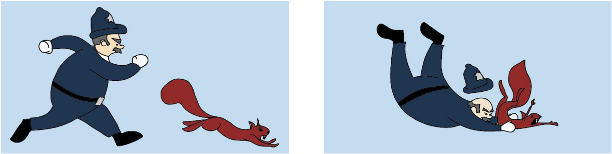
Following Abusch (2012) and Abusch & Rooth (2017), Maier & Bimpikou (2019) introduce PicDRT, a rather minimal extension of standard DRT (Kamp, 1981) for wordless comics. In a nutshell: the PicDRT construction algorithm adds the current panel in a new, labeled 'picture condition', identifying the 'regions of interest' (corresponding to salient individuals), and tagging these regions with fresh discourse referents. A postsemantic, pragmatic enrichment module takes care of linking some of the newly introduced discourse referents up with already established ones.

To illustrate, I reuse a minimal example from Maier & Bimpikou (2019), the visual analogue of a two-sentence discourse from a well-known introduction to DRT (Geurts, 1999):

(6) a. A policeman was chasing a squirrel. He caught it.

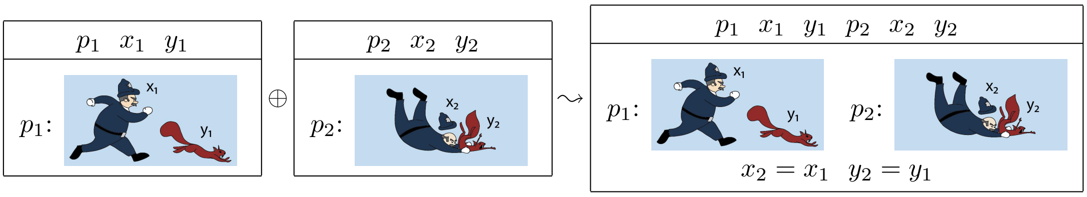
[2] Obviously, none of these or any other pictures in the paper are mine. They are panel fragments taken from googled images. These two *Donald Duck* panels are from an image published at https://www.nrc.nl/nieuws/2015/12/10/donald-duck-verrijkt-de-taal-1565843-a1298767.

b.



The first picture is interpreted as adding three discourse referents: p_1 representing the viewpoint of the entire picture; and x_1 and y_1 representing the individuals corresponding to the two salient picture regions (which we assume can be identified by the preliminary DRS construction algorithm, i.e. at the level of syntax/semantics, before pragmatic reasoning). The second picture is treated similarly, introducing three new discourse referents. After merging the two PicDRSs we pragmatically enrich the representation by inferring that the individual depicted in region x_1 is (probably) the same as that depicted in x_2, on the basis of certain plausible assumptions about similarity, world-knowledge, temporal progression between panels and other aspects of coherent story-telling.

(7)



Pragmatic enrichment need not stop here. Typical readers will view the picture sequence not just as depicting two states of affairs in which two entities reoccur, but add that the one is a policeman, and the second a squirrel, etc. We can add these defeasible pragmatic inferences as additional DRS conditions on x_1 and y_1 in the PicDRS. We return tho descriptive enrichment below, after we've introduced events.[3]

To interpret PicDRS boxes in a model we use an individual assignment function f to verify the regular DRS conditions, and a viewpoint assignment v to verify the pictorial conditions.

(8) a. model: $M = \langle D, W, I, V, \pi \rangle$ (where D is domain of individuals, W set of worlds, I and interpretation function, V a set of viewpoints, and π a projection function)

b. viewpoint assignment: $v :\subset \{p_1, p_2, \ldots\} \rightarrow V$

c. individual assignment: $f :\subset \{x_1, x_2, y_1, \ldots\} \rightarrow D$

The semantic definition of truth for PicDRSs combines a standard DRT semantics for descriptive conditions involving individual discourse referents, with a projective picture semantics for the picture conditions. Since the individual discourse referents thus refer to individuals via f, but also correspond to picture regions that refer to space-time regions via v (given π), we must make sure that f and v are properly 'aligned'. Technically, this last step requires that we extend the projection function π to map not only worlds (seen from a viewpoint) to pictures, but also individuals in those worlds (seen from a viewpoint) to picture regions. For concreteness we apply here the semantics to the final, pragmatically enriched example DRS in (7).

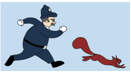
(9) $[\![(7)]\!]^{w,v,f} = 1$ iff there is a verifying embedding $f' \supset f$ with $Dom(f') = \{x_1, y_1, x_2, y_2\}$ and a viewpoint assignment $v' \supset v$ with $Dom(v') = \{p_1, p_2\}$ such that:

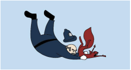
a. f' verifies the descriptive conditions:

[3] Cf. Wildfeuer (2019) for a different take on the DRT modeling of pictorial narratives, without projective picture conditions and without a distinction between semantic DRS construction and post-semantic enrichment.

(i) $f'(x_2) = f'(x_1)$
(ii) $f'(y_2) = f'(y_1)$

b. v' verifies the pictorial conditions:

(i) $\pi(w, v'(p_1)) =$ 

(ii) $\pi(w, v'(p_2)) =$ 

c. f' and v' are aligned:

(i) $\pi(f'(x_1), w, v'(p_1)) =$

(i.e. the policeman-region in the picture is a projective depiction of the policeman-entity $f'(x_1)$ in w, from the viewpoint associated with the picture)

(ii) $\pi(f'(y_1), w, v'(p_1)) =$

(iii) $\pi(f'(x_2), w, v'(p_2)) =$

(iv) $\pi(f'(y_2), w, v'(p_2)) =$

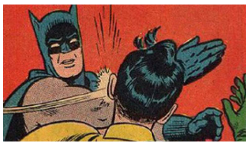
4 Picturing events

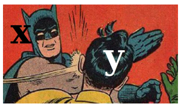
Speech and thought bubbles are readily thought of as the visual language analogue of quotation marks in written language (Saraceni, 2003; Cohn, 2013). To make this intuition precise I propose to extend PicDRT with some Davidsonian event semantics and then apply the powerful event-modification semantics of quotation (and reporting constructions more generally, see Kratzer 2006; Davidson 2015; Maier 2017).

Intuitively, we tend to read sequences of panels as representing sequences of events – in fact, that is what makes it a narrative. Panels should thus be understood as depicting (and introducing appropriate discourse referents for) not just individuals, but also events. Panels thereby become more like full-fledged utterances, which likewise introduce event discourse referents, though there it happens compositionally, through the lexical semantics of verbs. One further benefit of introducing events in visual narrative panels is that it's a first step toward applying a more general theory of discourse structure like SDRT, but for reasons of space I leave that general extension for another occasion.

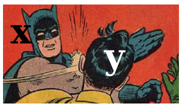
It is tempting to simply treat each panel as introducing a single (main) event discourse referent. But it's not hard to find counterexamples: panel sequences depicting stages of a single event, or depicting stative scenes (without temporal progression), or single panels depicting multiple (non-simultaneous) events (McCloud, 1993). If we look at comics with lots of dialogue and/or action we quickly find single panels with multiple speech and thought balloons attaching to different characters (who may also be moving around or experiencing emotions at the same time, perhaps indicated by further event modifiers as in (5)). One drastic option would be to leave the introduction of event discourse referents entirely up to pragmatic enrichment. Semantically speaking, the picture depicts a scene in a world, with a number of salient individuals, and on the basis of world knowledge we may infer that some particular type of event is probably happening. I opt for a middle position between requiring a general semantic one-event-per-panel rule and a leave-it-to-pragmatics strategy.

I propose to link the introduction of event discourse referents to the introduction of individuals. Each salient individual, as identified by the DRS construction algorithm, by stipulation, participates in some eventuality. Concretely, let's say that the DRS construction algorithm adds with each individual discourse referent x a new eventuality discourse referent e, and a condition stating that x participates in e ($\text{partcpt}(x, e)$). On the basis of the picture, world-knowledge, context, and rationality assumptions we may then post-semantically infer exactly what thematic roles and what (kinds of) eventualities are depicted ($\text{agent}(e, x)$, $\text{chase}(e)$, $\text{slap}(e)$,$\ldots$), at the same time as we're adding descriptive properties of the individuals ($\text{policeman}(x_1)$, $\text{batman}(x)$), and adding anaphoric relations between old and new discourse referents in the case of a multipanel narrative ($x_2 = x_1$, $e' \prec e$, $\ldots$). In the example below we have tow characters, which leads the construction algorithm to introduce two associated event discourse referents. In this case, post-semantic pragmatic reasoning might equate the two events, leaving just one event, a slapping, with two individuals fulfilling different thematic roles (Batman the agent, Robin the patient).

(10)  $\rightsquigarrow$

$p \quad x \quad e \quad y \quad e'$
p : 
$\text{partcpt}(x, e) \quad \text{partcpt}(y, e')$

$\rightsquigarrow$

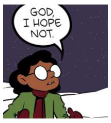
$p \quad x \quad e \quad y \quad e'$
p : 
$\text{partcpt}(x, e) \quad \text{partcpt}(y, e')$
$\text{agent}(e, x) \quad e' = e \quad \text{theme}(e', y) \quad \text{slap}(e) \quad \text{batman}(x) \ldots$

5 Bubbles

To deal with speech bubbles (and other symbolic enrichments exemplified in (5)) we'll need to decompose the panel into its (syntactic) constituents: the picture itself[4] (to be intepreted projectively, as part of a picture condition) and the bubble (to be interpreted decriptively, as a linguistic quotation). To preserve the connection between the speech bubble and the individual it 'points to' in the picture we use the discourse referent that the construction algorithm also generates for salient picture regions. From here on let's denote the first phase of interpretation, the generation of a DRS representation of a panel (or meaningful sign more generally), with $\langle\!\langle\ \rangle\!\rangle$ (because we like to reserve $[\![\ ]\!]$ for the second phase, i.e. the modeltheoretic interpretation of DRSs).[5]

[4] Note that once we disentangle the speech balloon and the picture we are left with a hole in the picture. The projective semantics needs a bit of adjustment to properly interpret holes as underspecified regions.

[5] The example is a fragment of a panel from https://www.smbc-comics.com/comic/complex

(11) $\langle\!\langle$ 

$\rangle\!\rangle =$

$p\quad x\quad e$

p:

$\text{partcpt}(x,e)$

$\oplus\ \langle\!\langle$

$\rangle\!\rangle(x)$

We analyze bubbles semantically, in the construction algorithm, as quotations, i.e. operators that take a linguistic input, a string of letters (written inside the bubble, as a kind of infix notation), to yield a property of individuals, viz. saying something of that form. This gives it the right type to apply to an individual discourse referent, as needed in (11).[6]

(12) a. $\langle\!\langle$ $\rangle\!\rangle = \lambda s\lambda y.\boxed{\begin{array}{c} e' \\ \hline \text{say}(e')\ \ \text{agent}(e',y) \\ \text{form}(e',s) \end{array}}$

b. $\langle\!\langle$

$\rangle\!\rangle = \langle\!\langle$ $\rangle\!\rangle(\ulcorner\text{God, I hope not}\urcorner) = \lambda y.\boxed{\begin{array}{c} e' \\ \hline \text{say}(e')\ \ \text{agent}(e',y) \\ \text{form}(e',\ulcorner\text{God, I hope not}\urcorner) \end{array}}$

We can now plug (12) into (11), to get the output of the construction algorithm:

(13)

$p\quad x\quad e\quad e'$

p:

$\text{partcpt}(x,e)$

$\text{say}(e')\ \ \text{agent}(e',x)$

$\text{form}(e',\ulcorner\text{God, I hope not}\urcorner)$

In words, the picture depicts an individual x who participates in some event e, and who is also the agent of a speech event e', which exhibits the linguistic form "God, I hope not".

Note that I've now analyzed the speech balloon as introducing a new event discourse referent e' of its own, rather than automatically linking to the default event associated with the individual (stipulated in 4). Of course, this leaves open the possibility of equating the two event variables post-semantically, pragmatically simplifying the semantic representation generated by the construction algorithm into one where what we see is an individual who is the agent of a particular speech event. The advantage of the current proposal is that it straightforwardly allows one panel to depict a single individual with various speech balloons originating from them, who is also participating in some salient non-linguistic activity.

Other bubble shapes conventionally encode other event properties. Cloud-like contours indicate thoughts, (replace 'say(e)' with 'think(e)' in (12a), see Maier (2017) on thought quotation more generally), bolded spiky contours indicate shouting, etc.

[6] For readability I freely apply lambda conversions in DRS construction derivations.

6 Beyond bubbles

The basic analysis of verbal bubbles provided here can be further extended in various directions. For instance, Cohn's (2013) non-verbal affixes, like the aforementioned motion lines, idea-lightbulbs, head-circling stars, or bulging vein anger symbols in (5), can be analyzed rather similarly. In the DRS construction phase, like bubbles they are isolated from the picture to be interpreted symbolically, as conventionally (not projectively) denoting properties of individuals, and then they are applied to the discourse referent associated with the region they were visually attached to.

Some of the symbolic embellishments in (5) are not obviously attached to a specific individual. Perhaps some sound effect descriptions are better interpreted as modifiers of events, and thus semantically applied to the event discourse referents, that are in turn introduced by individual depictions or other symbols. Action stars and fight clouds may just introduce events, only pragmatically linked to entities represented in the surrounding narrative (Cohn & Wittenberg, 2015). Further research is required to properly analyze the various of ways in which pictorial representations interact with symbolic enhancements.

As a final remark on thought bubbles, comics allow for the use of pictures inside bubbles to represent an agent's thoughts or other mental states iconically. We can model this with our event-based semantics by generalizing the predicate 'form(e, s)' to take a picture as its second argument (s). The idea is that thought events, imaginative projects, and similar mental processes can be either linguistic or visual. If s is a picture and e a thought event, 'form(e,s)' means that the thought is essentially a visual experience of the worlds that are projective mapped to the given thought bubble picture. Interestingly, the viewpoint of the thought bubble picture need not coincide with the *de se* center of the experience, leading to distinction parallel to that between free perception panels (Abusch & Rooth, 2017) and blended perception panels (Maier & Bimpikou, 2019).

References

Abusch, Dorit. 2012. Applying Discourse Semantics and Pragmatics to Co-reference in Picture Sequences. *Proceedings of Sinn und Bedeutung* 17.

Abusch, Dorit. 2015. Possible worlds semantics for pictures. In *The Blackwell Companion to Semantics*, .

Abusch, Dorit & Mats Rooth. 2017. The formal semantics of free perception in pictorial narratives. *Proceeding of the 21st Amsterdam Colloquium* .

Clark, Herbert & Richard Gerrig. 1990. Quotations as demonstrations. *Language* 66(4). 764–805.

Cohn, Neil. 2013. *The Visual Language of Comics: Introduction to the Structure and Cognition of Sequential Images.* A&C Black.

Cohn, Neil & Eva Wittenberg. 2015. Action starring narratives and events: Structure and inference in visual narrative comprehension. *Journal of Cognitive Psychology* 27(7). 812–828. doi:10.1080/20445911.2015.1051535.

Davidson, Kathryn. 2015. Quotation, demonstration, and iconicity. *Linguistics and Philosophy* 38(6). 477–520. doi:10.1007/s10988-015-9180-1.

Fernando, Tim. 2020. Pictorial narratives and temporal refinement. *Semantics and Linguistic Theory (SALT)* 29.

Geurts, Bart. 1999. *Presuppositions and Pronouns.* Amsterdam: Elsevier.

Giardino, Valeria & Gabriel Greenberg. 2015. Introduction: Varieties of Iconicity. *Review of Philosophy and Psychology* 6(1). 1–25. doi:10.1007/s13164-014-0210-7.

Goodman, Nelson. 1976. *Languages of Art: An Approach to a Theory of Symbols.* Hackett Publishing.

Greenberg, Gabriel. 2013. Beyond Resemblance. *Philosophical Review* 122(2). 215–287. doi:10.1215/00318108-1963716.

Henderson, Robert. 2016. A demonstration-based account of (pluractional) ideophones. *Semantics and Linguistic Theory* 26(0). 664–683. doi:10.3765/salt.v26i0.3786.

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.

Kaplan, David. 1989. Afterthoughts. In Joseph Almog, John Perry & Howard Wettstein (eds.), *Themes from Kaplan*, 564———. Oxford: Oxford University Press.

Kratzer, Angelika. 2006. Decomposing attitude verbs Handout. Honoring Anita Mittwoch on her 80th birthday. The Hebrew University of Jerusalem.

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.

Maier, Emar & Sofia Bimpikou. 2019. Shifting perspectives in pictorial narratives. *Sinn und Bedeutung* 23.

McCloud, Scott. 1993. *Understanding Comics.* Tundra Publishing.

Recanati, François. 2001. Open quotation. *Mind* 110(439). 637–687. doi:10.1093/mind/110.439.637.

Rooth, Mats & Dorit Abusch. 2017. Picture descriptions and centered content. *Proceedings of Sinn und Bedeutung* 21.

Saraceni, Mario. 2003. *The Language of Comics.* Psychology Press.

Schlenker, Philippe. 2018. Iconic pragmatics. *Natural Language & Linguistic Theory* 36(3). 877–936. doi:10.1007/s11049-017-9392-x.

Wildfeuer, Janina. 2019. The Inferential Semantics of ComicsPanels and Their Meanings. *Poetics Today* 40(2). 215–234. doi:10.1215/03335372-7298522.

Zimmermann, Thomas Ede. 2016. Painting and opacity. In Freitag et al. (ed.), *Von Rang und Namen: Philosophical Essays in Honour of Wolfgang Spohn*, 425–453. Mentis Verlag.