Tagging: Semantics at the Iconic/Symbolic Interface

Gabriel Greenberg

University of California, Los Angeles
gabriel.greenberg@gmail.com

Abstract

Tagging is the phenomenon in which regions of a picture, map, or diagram are annotated with words or other symbols, to provide descriptive information about a depicted object. The interpretive principles that govern tagged images are not well understood, due in part to the difficulty of integrating pictorial and linguistic semantic rules. Rather than directly combining these rules, I propose to use the framework of *perspectival feature maps* as an intermediary representation of content, in which the outputs of pictorial and linguistic interpretation may be assimilated. The result is a simple and compositional semantics for tagged images.

1 Tagging

Human communication is multi-modal. Spoken and signed words are accompanied by pictorial gestures and emotive facial expressions. Written words are enriched with illustrations, diagrams, and emoji. Newspaper articles come with photographs, and photographs come with captions. Maps are annotated with detailed geographic labels. And technical illustrations contain numerals, labels, call-out boxes, and more. In nearly every domain of human transaction, symbolic and iconic signs are integrated to efficiently express rich tapestries of information. Ultimately, the science of semantics must come to terms with this multi-modal outpouring.

We can loosely categorize multi-modal signs into three basic types. In **egalitarian** representations, icons and linguistic signs each express their own modality-specific content, which together contribute to a richer discourse content. Egalitarian representation are exemplified by sentential captions on photographs (Alikhani and Stone, 2019), by illustrated narratives and instructions (Alikhani and Stone, 2018b), and by many cases of coverbal gesture (Lascarides and Stone, 2009a,b).

In **language-dominant representation**, icons are used to enrich a linguistic expression; the interpretation of the iconic elements modulate the semantic contribution of the linguistic whole. Language-dominant multi-modality has been widely studied in recent years. Examples include pro-speech, co-speech, and post-speech gesture (Schlenker, 2018a; Tieu et al., 2017), iconic modulation of words (Schlenker, 2018a), and a wide variety of iconic enrichments in sign language (Schlenker, 2018b). See Schlenker (2019) for an overview of the state of the art.

The last category of multi-modal representation has received relatively little attention from semanticists, but is ubiquitous in human affairs. In **icon-dominant representation**, words, phrases, and non-linguistic symbols are used to enrich a picture, map, or diagram. The interpretation of the symbolic elements modulate the semantic contribution of the dominant image. **Tagging** is the phenomena in which symbolic **tags** are associated with specific subregions of the dominant icon.

Maps provide a vivid example of tagging, where names are inscribed throughout the map to indicate the location and identity of landmarks. Other cases include the descriptions, labels, and numerals featured in technical drawings and mass-media imagery, and the variable letters that are used to supplement mathematical and logical diagrams. Speech balloons in comics are

a genre-specific form tagging. And in, sign languages, classifier constructions involve the use of symbolic hand shapes to tag iconically presented paths and locations.

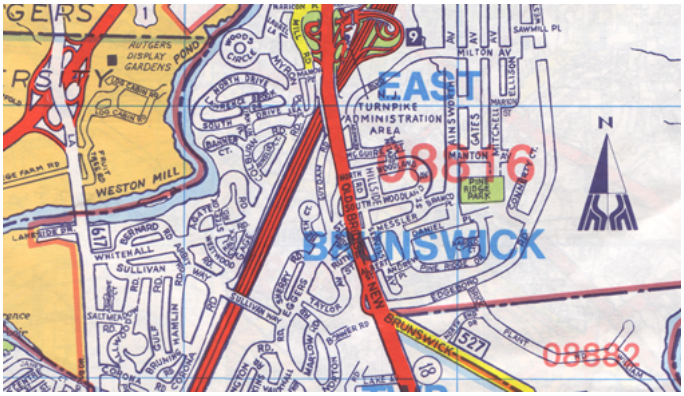
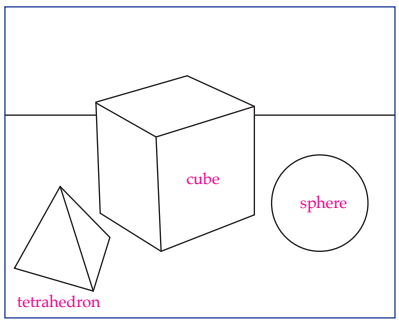


Figure 1: Tagging as the annotation of an icon with linguistic symbols.
Excerpt from from *Middlesex County Atlas* (2002), pgs. 32-33.

In this paper I'll focus on the paradigm case of **tagged images**: perspectival pictures enriched with linguistic tags. (I'll use "picture" and "image" interchangeably; I treat maps as a class of picture.) In tagged images, symbolic tags play the role of contributing identifying and descriptive information about particular *objects* (broadly construed) for which the picture provides pictorial information. An adequate semantics of tagging must address three central explanatory problems.

Problem 1: semantic significance of tag placement. The placement of tags within a printed images contribute to the accuracy-conditions of the whole tagged image by indicating which depicted objects are associated with the contents of the tags. For example, relative to a realistic scene, (1) is accurate. But swapping the position of "sphere" and "cube" results in (2) which is not accurate. In effect, tagging locates linguistically expressed properties within pictorial space. Understanding how this is done is the central challenge for a semantics of tagging.

(1)



(2)

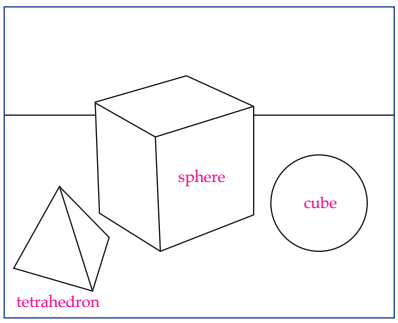


Figure 2: Relative to a realistic scene, image (1) is accurate, but (2) is not. They differ only with respect to the placement of "sphere" and "cube."

Problem 2: flexibility of tag placement. The ultimate semantic contribution of tags—to supply descriptive information about depicted objects– can be achieved through a variety of

expressive means. In Figure 3, for example, the tag "sphere" is associated with the image by placement *proximal* the part of the image it tags; "cube" is associated with a part of the image by placement *within* it; and "tetrahedron" is associated with the image through *line linking* (or *indication* in Alikhani and Stone 2018a, pg. 3555).

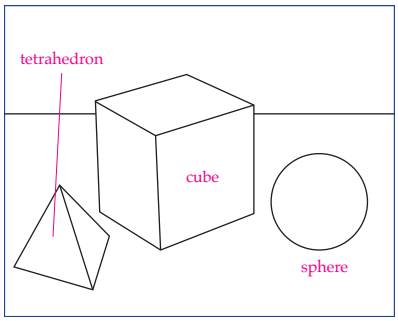


Figure 3: Tagging can be expressed through a variety of visual means: line linking, inclusion, and proximity. (Problem 2)

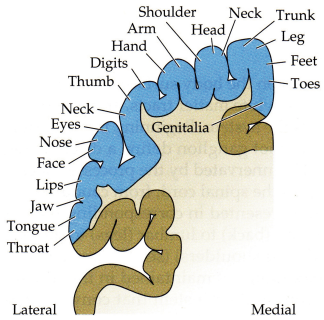


Figure 4: Tagging relations beyond predication in a diagram of the sensory cortex. (Problem 3) From Purves et al. (1997) *Neuroscience*, pg. 22.

Problem 3: variety of tagging relations. Nouns and names can both be tags, but because nouns and names denote objects in different semantic categories, the semantic significance of tagging must itself be allowed to vary. A picture of a person tagged with the name "Kiara" expresses the content that the person *is identical with* ⟦Kiara⟧; a picture of person tagged with the noun "professor" expresses the content that the person *has the property* ⟦professor⟧. More extreme variations are common. Consider the tag "08816" on a map; here the relevant relations is *has the zipcode*. In Figure 4, the relation *brain region X processes information from body region Y* is put to work in a standard depiction of the somatosensory cortex. As these examples show, the correct tagging relation cannot be determined simply as a matter of syntactic or semantic type, but must advert to contextual and discourse-sensitive constraints. Instead, I view tagging relations as a species of multi-modal **coherence relation** (Alikhani and Stone, 2018b), a structural link in discourse which functions to bind together independent discursive elements.

The prospect of a tagging semantics which answers Problems 1-3 presents us with a theoretical puzzle. On one hand, linguistic and pictorial elements demand radically different kinds of semantic analyses (Giardino and Greenberg, 2015; Schlenker, 2019). On the other hand, they cannot be entirely separated: to compute the content of a tagged image, one cannot simply divide it into a picture and a set of tags, compute their respective contents, and put them back together again. There would be no way of tracking which properties went with which depicted objects when they were recombined. In this paper, I'll recruit the theoretical apparatus of *feature maps* to serve as the nexus point where linguistic and pictorial information streams may come together. This approach will ultimately make it possible to formulate a simple and compositional semantics for tagged images.

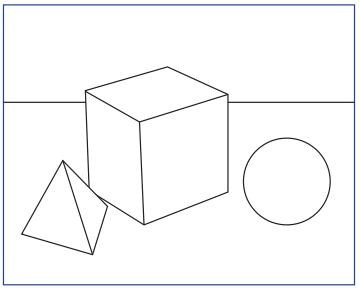
2 Syntax

The underlying syntactic structure of a tagged image can be divided into a **pure image** that is free from tags, a set of tags, and a set of **linking** relations that hold between regions of the pure image and pairs of tags and relation-symbols (cf. Alikhani and Stone, 2018b, pg. 2). On

this model, the location of a tag on the printed page is not itself part of syntax, but is a *signal* of a syntactic relation. Tags themselves have no location; they are associated with regions of the picture plane by abstract syntactic links. The rationale for this way of approaching tag placement will emerge shortly. Here the syntax of a tagged image stands to the image on the printed page roughly as a sentence's syntax stands to its phonology. I model this structure formally as follows, illustrated in Figure 5.

(3) A **tagged image** $T = \langle I, tag \rangle$, where:
  (i) I is a pure image;
  (ii) tag is a (partial) function from regions of I to pairs $\langle s, r \rangle$ where s is a tag-symbol, and r is a relation-symbol.

(4)



(5)

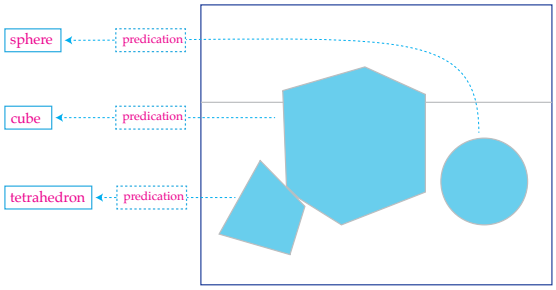


Figure 5: The syntax of a tagged image: (4) a pure image; (5) a set of linking relations between picture regions and pairs of tags and relation-symbols.

The structure of the pure image itself can be understood as a 2D plane segment where regions are associated with colors. (A more complete account might include lines and line-types, textures, and color regions.) Formally:

(6) A **pure image** $I = \langle P, d, color \rangle$ where:
  (i) P is a set of points;
  (ii) d is a Euclidean metric over P which defines a 2D rectilinear space;
  (iii) $color$ is a (total) function from points or small regions of P to colors (or values).

Meanwhile the linguistic constituents of tagged images are sub-clausal phrases that include names, numerals, nouns, adjectives, as well as definite and indefinite descriptions. In this paper, I'll set aside indexical sentences, like "you are here," which can also play a tagging-like role. Non-linguistic tags, as in a map, include a variety of specialized symbols which may be listed in a legend or conventional for a type of discourse.

Tags themselves are associated with regions of the picture plane by abstract structural links.[1] I'll assume that the regions in question are normally contiguous and correspond to psychologically natural segmentations of the graphical space. The links which connect tags to regions need not be explicitly marked on the printed page, but they are nevertheless signaled through a variety of means. Here a range of defeasible conventions may be invoked to indicate structural links:

[1] Such links are a sub-segmental variant of the text-to-image links posited by Alikhani and Stone (2018b). The semantic function of image sub-regions is anticipated in Abusch's analysis of visual co-reference (Abusch 2012, §3-5; Abusch 2015, §4). I'll assume here that the shapes of the regions in question are perfectly definite, though this is certainly an idealization in many cases.

1. **Inclusion:** a symbol tags the region it is located inside of.
2. **Proximity:** a symbol tags the region whose perimeter it is closest to.
3. **Alignment:** a symbol tags the region with an edge it is spatially aligned with.
4. **Line Linking:** a symbols tags the regions it is linked to by a line or arrow.

Figure 3 features line linking, proximity, and inclusion. Alignment can be seen in Figure 1, in the placement of road and river names.

These conventions often come into conflict. If a tag is *proximal* to one region, then it is necessarily *included* in a different region; which convention applies? The adjudication of these conflicts is sometimes guided by strict selection rules; for example, line-linking always trumps proximity and inclusion. But choosing between proximity and inclusion seems to be more open-ended, informed by spatial cues (e.g. degree of proximity), by semantic match between a picture region and the tag (e.g. "sphere" probably goes with the picture of the sphere), and by known design constraints (e.g. a long word like "tetrahedron" cannot be included in the region it tags). Determining which region is tagged by a symbol is a complex problem that may involve visual cognition, world-knowledge, and general purpose reasoning, in addition to specific conventions.

The proposed solution to Problem 2 is that the variety of expressions of tagging on the page all correspond to a single underlying relation of linking at the syntactic level. Inclusion, proximity, and line-linking are all signals of the underlying syntax, but they are not part of it. This analysis reflects a choice about where to draw the syntax/semantics boundary for the interpretation of tagged images. It divides interpretation into the pre-semantic process of disambiguating tagging links between symbols and regions, on one hand, and the semantic process of computing their meanings, on the other. Part of the rationale for this division of labor is that the two processes demand different kinds of cognitive capacities. The pre-semantic process of disambiguation requires defeasible reasoning and world-knowledge. The semantic process, by contrast, follows a set of narrowly defined interpretive rules, as I'll show. This bifurcation reflects the traditional view in linguistics and philosophy of language, which allows that general purpose reasoning may be enlisted in syntactic disambiguation, whereas semantics follows monotonic and compositional rules.

The final ingredient in the syntax of tagged images addresses the variety of tagging relations from Problem 3. I propose a set of **relation-symbols** which are explicit in the syntax, but implicit on the printed page. Each link between a symbol and a region is associated with one such symbol. Formally, I treat the *tag* function as a mapping from picture regions to pairs of tag symbols and relation-symbols. I'll represent *identity* and *predication*, the two most common tagging relations, by the relation-symbols as `id` and `pred` respectively.[2]

3 Content

What kind of contents are expressed by tagged images? Because tagging ultimately involves the location of linguistic information within pictorial space, we should model the contents of tagged images after the contents of pure images, rather than those for words or sentences. A popular approach to the contents of pictures understands them as sets of viewpoint-centered

[2]Despite variation, there are constraints on how tagging relations may be expressed. Identity and predication appear to be defaults. Other relations are inferred when these defaults are incoherent or otherwise ruled out in context. A further constraint is typographic consistency: tags with the same typographic features are expected to encode the same tagging relation. In Figure 4, for example, tags presented with line linking express the relation *processes information from*, while the two tags "Lateral" and "Medial" in the bottom corners, which use proximity, express predication. Ultimately these factors must be considered within the context of discourse coherence theory more generally.

worlds (Blumson, 2009; Abusch, 2015). But a centered-worlds approach is an awkward fit for the object-oriented semantics of tagging. Tags are associated with *objects*, not entire scenes. Once the set of centered-worlds associated with a pure image is fixed, there is no way of going back into the set and introducing the semantic contribution of the tags, without re-computing the content of the image.

Instead, I propose to use a level of structured content which is intermediate between syntax and accuracy conditions. We can define a simple semantics which separately maps visual and symbolic contents into this intermediate level, which in turn is subject to a general definition of accuracy. The intermediate level is a type of **feature map**, a 2D array whose regions are associated with objects and properties. Feature maps preserve the visual structure of the underlying image while trading the syntactic constituents of the picture for the semantic elements they express. In Greenberg (2019b) I develop a model of pictorial contents as **perspectival feature maps** (PFMs), a type of feature map where each point in the array is associated with a viewpoint-centered direction. Feature maps provides an intuitive interface between the pictorial sign and the background projection semantics, and are straightforwardly extended to incorporate the contents of tagged images.

A perspectival feature map is a two-dimensional array where each point in the array is associated with a viewpoint-centric direction in three-dimensional space; and regions of the array are associated with clusters of objects and properties.[3]

(7) A **perspectival feature map** $M = \langle Field, Direction, Cluster \rangle$ where:
- (i) *Field* is a two-dimensional array;
- (ii) *Direction* is a total function from points in *Field* to 3D directions which satisfy a viewpoint condition.
- (iii) *Cluster* is a (partial) function from regions of *Field* to feature-clusters.

A **feature cluster** is a sequence $\langle o, G_1, ..., G_n \rangle$ where o is an object and $G_1, ..., G_n$ are properties. If a picture express a feature map as content, then the objects of the map's feature clusters correspond to the singular contents of the picture. These are the objects it depicts. The properties of the feature cluster correspond to the attributive contents of the picture. These are the properties it depicts its objects as having (Greenberg, 2018). The structure of PFMs is illustrated in the two figures below.

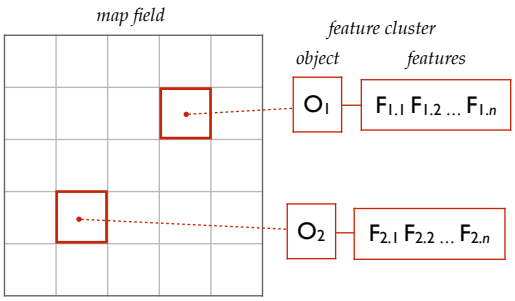


Figure 6: Feature map with feature clusters.

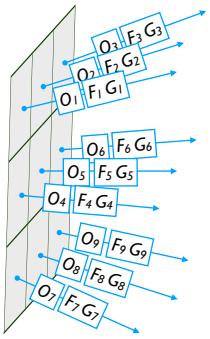


Figure 7: Perspectival feature map with feature clusters and viewpoint-centric directions.

[3] A more complete definition of feature maps would allow for variations in acuity, the representation of relations, the expression of more than one represented object per region (Greenberg, 2019b).

A perspectival feature map can be thought of as a kind of *directional space*— a space whose "dimensions" are directions emanating from a viewpoint, and whose constituents are the objects and properties laid out in that space. It locates each of the objects in its feature clusters in a given direction, and attributes to each its associated properties. A PFM is accurate relative to a world w and viewpoint v if and only if the attributions it makes are correct, when it is fixed to the location of v within w. In the definition below, given a PFM and its associated $Cluster$ function, let $object()$ be a function from a region r to the object of $Cluster(r)$, and $properties()$ to the set of properties in $Cluster(r)$.

(8) A perspectival feature map $M = \langle Field, Direction, Cluster\rangle$
is **accurate at** a world w and viewpoint v iff for every region $r \in dom(Cluster)$:
(i) $object(r)$ is located in $Direction(r)$ from v in w;
(ii) $object(r)$ realizes each $F \in properties(r)$ in w.

PFMs are a natural choice for the content of tagged images, because they easily accommodate the accumulation of features associated with different regions of the visual field. For pure images, the constituents of the feature clusters are, to a first approximation, entirely visual properties. For tagged images, the content of a tag is simply construed as yet another property which is added to the feature cluster of a depicted object. Thus the word "cube", used as a tag, expresses the property *cube*. Since the properties expressed by tags enter into feature clusters in the feature map, they automatically inherit the spatial significance of the structure of the map itself. Thus, if the property *cube* is associated with an object o in the feature map, the final accuracy conditions will simultaneously attribute *cube* to o and attribute a specific viewpoint-relative direction to o. The result is that tagged properties are projected out through pictorial space. In this way, the PFM framework allows us to directly co-index tagged properties with visually depicted properties within a single semantic structure. And this is ultimately what is required to capture the distinctive semantic contribution of tagging.

4 Semantics

The central challenge for a semantics of tagged images is the problem of integrating the deeply divergent semantic rules for words and pictures. In previous work I've advocated a projection semantics for pictures, where a picture is mapped to the set of centered-worlds from which it can be projected (Greenberg, 2019a). But this approach is not easily extended to tagging, since there is no way to locate the properties expressed by tags within the worlds expressed by the pure image, without drawing on the projective rule which defined the pure image content in the first place. Instead, the interpretation of tags would have to be integrated into the definition of projection from the start. The resulting theory could be systematic, but it would not be compositional, since the content of the pure image would not be computed independently of the contents of the tags.[4] On this approach, one can't simply use one's semantics for pure images "out of the box."

To get at the semantics of tagging, we must find a way to integrate the interpretation of the tags with the interpretation of the pure picture. It is here that the feature map framework comes into its own, for PFMs are naturally suited to the task of coordinating diverse streams of object-oriented information within a unified visual frame. My strategy is to exploit the parallelism between tagged image syntax and feature map structure: as each tag is associated

[4]Compare the theory of indexing in Abusch (2012, 2015), which integrates object-coreference with rules of projection.

with a region, it contributes a property to the feature cluster expressed by that region. The resulting account smoothly assimilates the content of the pure image and those of the tags.

By stating the semantic rules as mappings from syntactic elements to elements of a feature map, the interpretation of a complete tagged image can be neatly divided into three sub-problems: (i) the semantics for pure images; and (ii) the semantics for the tags themselves; and (iii) the semantics of the region-to-tag links. The resulting semantics for tagged images is compositional, in the sense that the content of the whole is a function of the content of each part (the pure image, the tags, and the relation symbols), and the way they are put together (the linking relations).

Each sub-problem is governed by distinct interpretive rules. The linguistic expressions in a tagged image are interpreted relative to a language. So too, pictures are governed by **systems of depiction**, the pictorial analogues of languages. The pictures of platonic solids used in this paper belong to a simple system of black and white line drawing. But systems vary in their treatment of line, color, shading, stylization, and more. Tagged images themselves belong to **hybrid systems**: the combination of system of depiction and a language. Where L is a language and D is a system of depiction, a tagged image belongs to the hybrid system L/D.

(9) **Tagging Semantics**
Given a hybrid system L/D, for any tagged image $T = \langle I, tag \rangle$ in L/D, and context c:
$[\![T]\!]_{L/D,c}$ = the minimal feature map $M = \langle Field, Direction, Cluster \rangle$ such that:
(i) **Congruence:**
there is a unique $f : P \mapsto Field$ such that I and $Field$ are congruent wrt to f;
(ii) **Pictorial Semantics:** $[\![I]\!]_{D,c} \sqsubseteq M$;
(iii) **Tagging:**
$\forall r \in dom(tag)$: where $tag(r) = \langle S, R \rangle : [\![R]\!]_{L/D}([\![S]\!]_L) \in properties(f(r))$.

Clause (i) serves the same function as before, imposing congruency between regions of the image surface and feature map regions. Clause (ii) introduces the semantic contribution of the pure image, presupposing something like the semantics specified above. The "$\sqsubseteq$" relation is the part-hood relation for feature maps.

Clause (iii) states that, for every tagged region r in the image, a corresponding property should be added to the feature cluster which is expressed by r in the feature map. The property in question is not simply $[\![S]\!]_L$, the content of the tagged symbol, but rather $[\![R]\!]_{L/D}([\![S]\!]_L)$, the content of the tagged symbol as it is modulated by the content of the relevant tagging relation, as required by Problem 3. The denotations for two most common relation-symbols, "`id`" (*identity*) and "`pred`" (*predication*) are:

(10) $[\![\texttt{id}]\!]_{L/D} = \lambda x \lambda y . x = y$

(11) $[\![\texttt{pred}]\!]_{L/D} = \lambda F . F$

This semantics has the desired effect of allowing the content of symbolic tags to enter into the content of the tagged image at specific, object-dependent locations in pictorial space. The resulting analysis provides a satisfactory account of Problem 1, the semantic contribution of tags to accuracy conditions. To see this, recall the accurate image (1) and inaccurate image (2) from Figure 2, which differed only in the placement of the tags "cube" and "sphere". Suppose that, in (1), "cube" is linked by predication to region r_1. And assume that $[\![\text{cube}]\!]_L$ = the property *cube*. By clause (iii) from tagging semantics, the "cube" tag imposes the following condition on the resulting feature map:

(12) a. $[\![\texttt{pred}]\!]_{L/D}([\![\text{cube}]\!]_L) \in properties(f(r_1)) \Leftrightarrow$

b. $cube \in properties(f(r_1))$

By the definition of accuracy for feature maps, it follows that the tagged image is accurate at a centered-world $\langle w, v \rangle$ only if:

(13) a. $object(f(r_1))$ is located in $Direction(f(r_1))$ from v in w;
b. $object(f(r_1))$ realizes the property $cube$ in w.

The inaccurate image (2) is exactly like the accurate (1), except that "cube" is associated with r_2, rather than r_1. When r_2 is substituted for r_1 in the accuracy conditions above, the picture locates a cube in a different direction within pictorial space. As a result, (1) and (2) express different accuracy conditions.

I turn next to pictorial semantics, the semantics governing pure images. Pictorial semantics is itself a multi-faceted problem, where vision, convention, and context all play a role in determining meaning (Kulvicki, 2006; Greenberg, 2018). I'll assume that, given a system of depiction D, image I, and context c, $[\![I]\!]_{D,c}$ is a PFM (Greenberg, 2019a). Context determines the singular content of a PFM by associating regions of I with objects. Systems of depiction, in turn, determine part of the attributive content of a PFM by associating regions with directions and basic properties. Further stages of visual processing contribute additional visual features like depth, 3D shape, and category. The semantics sketched below shows how these different interpretive vectors can be brought together:

(14) **Pictorial Semantics**
Given a system of depiction D, for any image $I = \langle P, d, color \rangle$ in D, and context c:
$[\![I]\!]_{D,c}$ = the minimal feature map $M = \langle Field, Direction, Cluster \rangle$ such that:
(i) **Congruence:**
there is a unique $f : P \mapsto Field$ such that I and $Field$ are congruent wrt f;
(ii) **Reference:**
$\forall r \subseteq P:$ if $r \in dom(ref_c)$, then $ref_c(r) = object(f(r))$;
(iii) **System of depiction:**
there is a viewpoint $v = vp_c(Field)$ such that $\forall p \in P:$
(a) **Projection condition:**
$Direction(f(p))$ is co-directional with $projection_D(v, f(p))$;
(b) **Marking condition:**
$marking_D(color(p)) \in properties(f(r))$;
(iv) **System of vision:**
$V_f(P) \sqsubseteq M$;

Clause (i) imposes a spatial congruency f between regions of the image surface and feature map regions, which is used to preserved consistency between the other clauses. Where r is an image region, $f(r)$ is the corresponding map region. Two 2D fields are **congruent** just in case there is a a metric isomorphism between them that preserves up/down and front/back orientations. (I've suppressed orientation vectors in the definitions of images and map fields here.) Clause (ii) accounts for the singular content of an image, as determined by context, which in turn reflects artist's intentions and the causal history of thte picture's production. Contextual reference is modeled as a function ref_c, part of a context c, which associates (some) regions of I with objects.

Clause (iii) specifies the contribution of the system of depiction within a projection semantics (Greenberg, 2013; Abusch, 2015). In Greenberg (2019a), I've shown how a projection semantics can be translated into a feature map framework; in short, projection semantics implies that each

feature cluster in a picture's feature map be association with (a) a viewpoint-relative direction; and (b) a basic feature (such as *surface*, *edge*, or color). These are determined, in turn, by the **projection condition** and **marking condition** which characterize the system of depiction. Formally, vp_c fixes the position of the viewpoint, relative to he map field, within the 3-space of the feature map; $projection_D$ is a function from viewpoints and points on the map field to rays in the 3-space of the feature map; $marking_D$ is a function from colors in the picture to basic features. Clause (iv) is a catch-all for the contribution of visual computation to pictorial content.

Drawing upon the feature map analysis of visual content, I have sketched a theory of the syntax, content, and semantics of tagged images. While the account leaves significant issues unresolved, I hope the general analytical strategy I've pursued here is flexible enough to extend to other kinds of tagging in maps, comics, sign language, and beyond.

References

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

Dorit Abusch. Possible worlds semantics for pictures. In Lisa Mathewson, Cécile Meier, Hotze Pullman, and Thomas Ede Zimmermann, editors, *Blackwell Companion to Semantics*. Wiley, New York, 2015. Forthcoming.

Malihe Alikhani and Matthew Stone. Arrows are the verbs of diagrams. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3552–3563, 2018a.

Malihe Alikhani and Matthew Stone. Exploring coherence in visual explanations. In *2018 IEEE Conference on MIPR Processing and Retrieval (MIPR)*, pages 272–277.

Malihe Alikhani and Matthew Stone. "Caption" as a coherence relationimplications. In *Proceedings of the 2nd Workshop on Shortcomings in Vision and Language*, pages 58–67, 2019.

Ben Blumson. Pictures, perspective and possibility. *Philosophical Studies*, 149(2):2009.

Valeria Giardino and Gabriel Greenberg. Varieties of iconicity. *Review of Philosophy and Psychology*, 6(1):1–25, 2015.

Gabriel Greenberg. Beyond resemblance. *Philosophical Review*, 122(2):215–287, 2013.

Gabriel Greenberg. Content and target in pictorial representation. *Ergo*, 5(23), 2018.

Gabriel Greenberg. Semantics of pictorial space. Manuscript, 2019a.

Gabriel Greenberg. The structure of visual content. Manuscript, 2019b.

John Kulvicki. *On Images: Their Structure and Content*. Clarendon, Oxford, 2006.

Alex Lascarides and Matthew Stone. Discourse coherence and gesture interpretation. *Gesture*, 9(2):147–180, 2009a.

Alex Lascarides and Matthew Stone. A formal semantic analysis of gesture. *Journal of Semantics*, 26(4):393–449, 2009b.

Philippe Schlenker. Iconic pragmatics. *Natural Language & Linguistic Theory*, 36(3):877–936, 2018a.

Philippe Schlenker. Visible meaning: Sign language and the foundations of semantics. *Theoretical Linguistics*, 44(3-4):123–208, 2018b.

Philippe Schlenker. What is super semantics? *Philosophical Perspectives*, 2019.

Lyn Tieu, Robert Pasternak, Philippe Schlenker, and Emmanuel Chemla. Co-speech gesture projection: Evidence from truth-value judgment and picture selection tasks. *Glossa: a journal of general linguistics*, 2(1), 2017.