


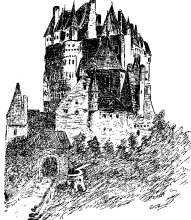
Indexing Across Media

Mats Rooth and Dorit Abusch

Cornell University
Ithaca, New York, USA
mr249@cornell.edu da45@cornell.edu

1 Introduction

Indexing across pictures and language is illustrated in (1). In the first example, the pronoun in the sentence somehow picks up a discourse referent set up by the picture, and the picture and the sentence jointly put constraints on the same individual in a described situation. In (1b) a nominal phrase functioning as a title or caption gives information about an individual depicted in the picture. This paper analyzes indexing in such examples, starting from a dynamic semantics for indexing in pictorial narratives. The current section reviews the semantic framework. The basic analysis of indexing across media is laid out in Section 2, using a setup involving a formal language and interpretation for it. Section 3 looks at data where in a combination of a picture and some language, the linguistic part is a nominal (such as a title) rather than a sentence. Here a constraint on interpretation is observed, which in the theory is enforced in the syntax of discourse representations. Section 4 looks at data involving definite reference and quantification. Section 5 points out purely linguistic data that are analogous to the data from Section 3. Section 6 wraps up.¹

- (1) a.  b. 
- He's a sailor. A castle owned by a duke
- Navy sailor
drinking coffee.
Openclipart.
- Castle on a hill.
Public Domain
Vectors.

We assume the semantics for pictures and indexing in pictorial narratives employed in previous work (Abusch, 2012, 2014, to appear; Abusch & Rooth, 2017; Rooth & Abusch, 2018; Maier & Bimpikou, 2019). The framework is reviewed briefly here.² A propositional semantics for pictures is based on geometric projection. The basis is a projection function π that maps a world and a viewpoint to a two-dimensional picture, using a mathematical, computational, and/or physically realized procedure such as perspectival projection or orthographic projection. A viewpoint is analogous to a camera position, or the station point in the classical theory of perspective. Where w is a world and v is a viewpoint, the function value $\pi(w, v)$ is the picture that is projected from world w as observed from viewpoint v . Propositional semantic values are then obtained by inverting projection. The propositional semantic value of a given picture p is the set of worlds that project to p via π . There are a handful of independent arguments for employing viewpoint-centered semantics values, which are sets of pairs of a world and a

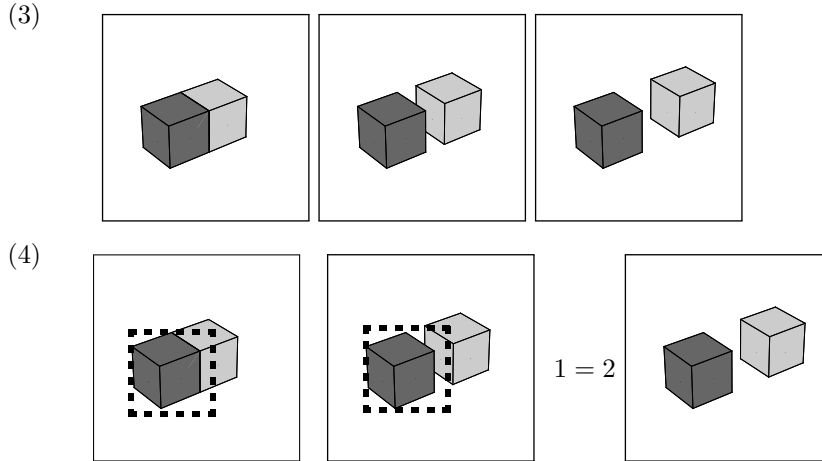
¹The images in the paper that are quoted from books and other sources are used for educational and critical purposes, and are property of their respective owners.

²Abusch (to appear) is a thorough review.

viewpoint.³ In this option, which is assumed here, the semantic value $\llbracket p \rrbracket$ of picture p is the set of pairs $\langle w, v \rangle$ such that w projects to p from viewpoint v , $\pi(w, v) = p$. This is recorded in (2a).⁴ (2b) is a variant where times are included in the model, the projection function has a time argument together with a world and a viewpoint, and the semantic value is a set of triples of a world, a time, and a viewpoint.

$$(2) \quad \text{a. } \llbracket p \rrbracket = \{ \langle w, v \rangle \mid \pi(w, v) = p \} \quad \text{b. } \llbracket p \rrbracket = \{ \langle w, t, v \rangle \mid \pi(w, t, v) = p \}$$

In the analysis of Abusch (2012, to appear) a picture or picture sequence is incremented syntactically with geometric areas, which introduce discourse referents. As an example, (3a) is a three-panel comic of two cubes moving apart. A basic semantics combines the semantics of the individual pictures with homomorphic temporal progression.⁵ This basic semantics (in a possible-worlds model with worlds and times) does not entail that in the described situation, the cube corresponding to the gray area in the first frame of the comic is the same as the cube corresponding to the gray area in the second frame (and so on). In order to express these understood identities, Abusch (2012) suggested incrementing pictures with *areas in the picture* that introduce discourse referents. In (4), there is a bounding box around the dark area in the first picture, and similarly in the second picture. These serve to introduce discourse referents for depicted individuals, in this case a discourse referent for the cube depicted in the first picture, and another discourse referent for the cube depicted in the second picture. The identity $1 = 2$ has the semantics of equality in the model, indicating that the individuals in the model that correspond to the two discourse referents are identical. Bounding boxes serving as proxies for individuals are used in machine learning databases and algorithms. For instance, (5) is an image from the Pascal VOC dataset with a picture of a bus and a bounding box for the bus (Everingham et al., 2012).



³These arguments include ones based on the semantics of discourse referents (as here), accounting for Necker ambiguities, and the use of perspectival phrases such as *in front of* in sentences describing pictures. See Abusch (to appear) and Rooth & Abusch (2018).

⁴A model M and parameters A of the projection function can be added outside the brackets, $\llbracket p \rrbracket^{M,A} = \{ \langle w, v \rangle \mid w \in M^W \wedge \pi^A(w, v) = p \}$.

⁵Abusch (2014) discusses temporal progression in visual narratives.

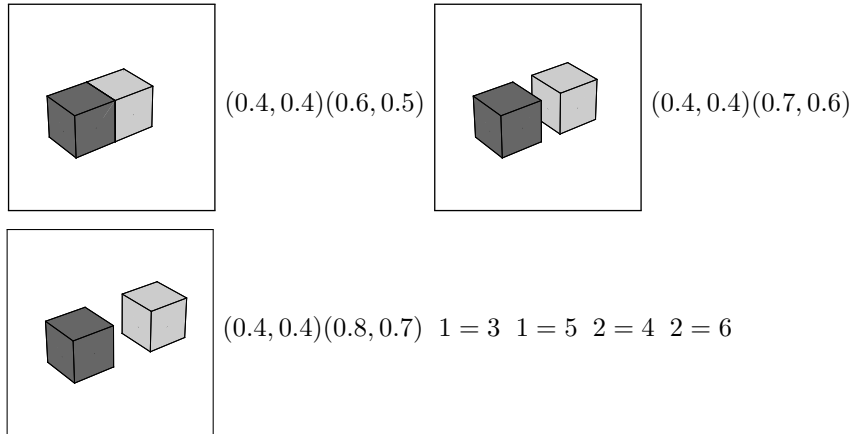
(5)



Pascal VOC dataset.

For simplicity, in this paper we use points in the area of the picture as geometric discourse referents. With the assumption that the first picture in (3) is a unit square, the pair $(0.4, 0.4)$ measures 0.4 along the horizontal axis and 0.4 along the vertical axis to a point within the dark gray area. The pair $(0.6, 0.5)$ measures 0.6 along the horizontal axis and 0.5 along the vertical axis to a point within the light gray area. (6) is a version of (3) that includes geometric discourse referents for all the depicted cubes. At the end there are equalities that use a recency convention. The equality $1 = 3$ equates the most recently introduced discourse referent with the ante-penultimate one. This has the effect of equating the light cube in the final picture with the light cube in the middle picture.

(6)



Something like (6) is a formula of a formal language with a defined syntax, and a semantics that is stated in type theory and possible worlds semantics. Abusch (to appear) formalized the semantics inductively, using a format similar to (7), where a world (variable w), a time (variable t), a viewpoint (variable v) and a string of individuals (here x_1x_2) satisfy a formula. v is the viewpoint for the last picture in the formula.

$$(7) \quad w, t, v, x_1, x_2 \models \left[\text{cube} \right] (0.4, 0.4)(0.6, 0.5)$$

The point of memorizing the viewpoint for the last picture is that this viewpoint is used in the semantics of discourse referents. Given a viewpoint v (understood as the viewpoint for the last picture), and a point d , understood as a point in the two-dimensional area of the last picture, v and d are used to pick out an object by tracing a directed line from v through the point d in the picture plane to the point where it intersects an object. An object that

witnesses the discourse referent is one that the directed line from v through d intersects before it intersects any other object. We write this condition as $\bar{\pi}(w, t, v, d, x)$, or when time is not being considered, as $\bar{\pi}(w, v, d, x)$.⁶

Let P be a visual narrative like (6), consisting of a sequence of pictures, with interleaved discourse referents and equalities between discourse referents. By collecting up the tuples that satisfy P , we obtain a semantic value for P , which is a set where each element is a tuple of a world, a time, a viewpoint, and witnesses for discourse referents. This is recorded in (8).

$$(8) \quad \llbracket P \rrbracket = \{ \langle w, t, v, x_1, \dots, x_n \rangle \mid w, t, v, x_1, \dots, x_n \models P \}$$

This is a set of cases in the sense of Lewis (1975). Lewis introduced case semantics to theorize about indefinite descriptions and anaphora in sentences with adverbs of quantification, such as the Murphy's law examples (9). He showed that by assuming a case semantics for the two clauses in such sentences (i.e. the if-clause and the main clause) it is possible to arrive at a semantics for the whole compositionally.

- (9) a. If you drop an unbreakable object, it always lands on something more valuable.
 b. If two cars are driving in opposite directions on a long road with a one-way bridge, they always meet at the bridge.

To deal with sentences that have free indices (such as the main clauses in (9)), it is necessary to say that a syntactic unit denotes a set of cases *relative to a case*. Where X is the syntactic unit and c is a case, for this we use the notation $c\llbracket X \rrbracket$.⁷ We will always refer to cases that are of the form $wv\mathcal{O}$ or (when time is being ignored) $wv\mathcal{O}$. Here \mathcal{O} is a string of objects that witness discourse referents, w is a world, t is a time, and v is a viewpoint. (10) gives some semantic values in this notation.

- (10) a. $wv\mathcal{O}\llbracket [\text{he}_1 \text{ has a dog}] \rrbracket = \{ c \mid \exists x. c = wvx\mathcal{O} \wedge \mathbf{dog}(w, x) \wedge \mathbf{have}(w, \mathcal{O}[1], x) \}$
 b. $wv\mathcal{O}\llbracket [1 = 2] \rrbracket = \{ c \mid c = wv\mathcal{O} \wedge \mathcal{O}[1] = \mathcal{O}[2] \}$

c. $wv\mathcal{O} \left[\text{img} \right] = \left\{ c \mid \exists v'. c = wv'\mathcal{O} \wedge \pi(w, v') = \left[\text{img} \right] \right\}$

2 A Basic Analysis

We have seen that the semantics value of an enriched pictorial narrative (as formulated in Section 1) is the same kind of formal object as the semantics of a sentence of English containing indefinite descriptions and pronouns. In Abusch (2012, to appear); Abusch & Rooth (2017); Rooth & Abusch (2018), this is used to give an analysis of indexing in pictorial narratives, and analyses of additional phenomena, using the toolkit of dynamic natural language semantics. Here we observe that, once we move to the semantics, there is no difference between indexing within a medium and indexing across media. An index that is set up within a pictorial narrative

⁶There are questions about the optimal formulation geometric discourse referents (e.g. points vs. bounding boxes) and the optimal definition of what objects correspond to them. For instance, if the individuals in the model have part-whole structure and we use points, there may be unwelcome multiplicity in the value of x . Consider a point d in the head-area of a depicted character. Let x be a person in the model, and let x_h be that person's head. If $\bar{\pi}(w, t, v, d, x_h)$ holds, then also $\bar{\pi}(w, t, v, d, x)$ holds. See for discussion Abusch (2014). Bounding boxes can be used to partially alleviate this problem. Ultimately though we are inclined to maintain that predications about the type of depicted objects are accommodated, e.g. $\mathbf{person}(w, x)$.

⁷Rooth (to appear) presents Lewis's semantics for indefinites and adverbs of quantification along these lines.

can be picked up later in the pictorial narrative. But equally, it can be picked up with a pronoun in a sentence of natural language.

Consider the left column in (11), which we think of as a scenario where a parent reading a picture book to a child points out a character in a picture, gives some information verbally, continues by pointing at (or touching) the dog in the next picture, and then adds some more verbal information.

(11) a.



His name is Dick.

He has a dog.



His name is Spot.

b.



$\langle \cdot, \cdot \rangle$

[his₁ name is Dick]

[he₁ has a dog]



$\langle \cdot, \cdot \rangle$

1 = 2

[his₁ name is Spot]

Images from William Gray, *Fun with Dick and Jane*, 1946.

The right column gives a counterpart in our formal language, where the finger-touching gestures are replaced by geometric discourse referents, and equalities between discourse referents are added. This formula has a linear structure with eight parts, which we name p_1 (a picture), a point d_2 introducing a discourse referent, a sentence s_3 containing a pronoun, a sentence s_4 containing a pronoun and an indefinite description, a picture p_5 , a point d_6 introducing a

discourse referent, an equality between discourse referents e_7 , and finally a sentence s_8 .⁸

The cross-medium narrative (11b) is to be interpreted in the uniform dynamic framework that was reviewed in Section 1. To simplify, in the current discussion we do not include times. (12) gives the semantics of the eight parts of the narrative. A picture p_i interpreted relative to $wv\mathcal{O}$ introduces a new viewpoint v' , and checks that the world from the viewpoint projects to the picture. \mathcal{O} is not incremented. Thus $wv\mathcal{O}\llbracket p_i \rrbracket = \{z | \exists v'. z = wv'\mathcal{O} \wedge \pi(w, v') = p_i\}$, where the new viewpoint is recorded in an output case $wv'\mathcal{O}$. A geometric discourse referent d_i non-deterministically chooses an object x , and checks the geometric constraint $\bar{\pi}(w, v, d_i, x)$ that relates the viewpoint v for the last picture, the point d_i , and the value x for the discourse referent. \mathcal{O} is incremented with x to form $x\mathcal{O}$. Thus $wv\mathcal{O}\llbracket d_i \rrbracket = \{z | \exists x. z = wvx\mathcal{O} \wedge \bar{\pi}(w, v, d_i, x)\}$. An equality $m = n$ is semantically a test that checks equality of $\mathcal{O}[m]$ and $\mathcal{O}[n]$, see (12g). The three sentences are given standard interpretations in dynamic semantics. Indexed pronouns look up their referents in \mathcal{O} , with indexing into \mathcal{O} following a recency convention. Thus the index 1 in sentence s_3 (“his₁ name is Dick”) gets the value $\mathcal{O}[1]$, and $\llbracket s_3 \rrbracket$ is a test which checks the name of $\mathcal{O}[1]$. The indefinite description in sentence s_4 introduces a new value x that is entered as $x\mathcal{O}$, which is constrained to be a dog in w , and to be possessed by $\mathcal{O}[1]$ in w .

- (12) a. $wv\mathcal{O}\llbracket p_1 \rrbracket = \{z | \exists v'. z = wv'\mathcal{O} \wedge \pi(w, v') = p_1\}$
 b. $wv\mathcal{O}\llbracket d_2 \rrbracket = \{z | \exists x. z = wvx\mathcal{O} \wedge \bar{\pi}(w, v, d_2, x)\}$
 c. $wv\mathcal{O}\llbracket s_3 \rrbracket = \{z | z = wv\mathcal{O} \wedge \mathbf{name}(w, \mathcal{O}[1], \text{“Dick”})\}$
 d. $wv\mathcal{O}\llbracket s_4 \rrbracket = \{z | \exists x. z = wvx\mathcal{O} \wedge \mathbf{dog}(w, x) \wedge \mathbf{have}(w, \mathcal{O}[1], x)\}$
 e. $wv\mathcal{O}\llbracket p_5 \rrbracket = \{z | \exists v'. z = wv'\mathcal{O} \wedge \pi(w, v') = p_5\}$
 f. $wv\mathcal{O}\llbracket d_6 \rrbracket = \{z | \exists x. z = wvx\mathcal{O} \wedge \bar{\pi}(w, v, d_6, x)\}$
 g. $wv\mathcal{O}\llbracket e_7 \rrbracket = wv\mathcal{O}\llbracket 1 = 2 \rrbracket = \{z | z = wv\mathcal{O} \wedge \mathcal{O}[1] = \mathcal{O}[2]\}$
 h. $wv\mathcal{O}\llbracket s_8 \rrbracket = \{z | z = wv\mathcal{O} \wedge \mathbf{name}(w, \mathcal{O}[1], \text{“Spot”})\}$

Thus the eight parts of the cross-medium narrative get interpretations in a uniform dynamic semantic framework. This immediately answers the question of how information from different media is combined: such information is combined in the way information in a single medium is combined in a dynamic framework, namely by dynamic conjunction. (13) is a formulation of dynamic conjunction in the current notation. Here x , y , and z are cases of the form $wv\mathcal{O}$, and the definition essentially expresses relation composition.

$$(13) \quad x\llbracket AB \rrbracket = \{z | \exists y [y \in x\llbracket A \rrbracket \wedge z \in y\llbracket B \rrbracket]\}$$

Conjoining the parts in (12) using dynamic conjunction results in (14) as the semantics of the cross-medium narrative (11b), relative to a null context wv consisting of a world and an (irrelevant) viewpoint.

⁸The phenomenon of touching pictures to set discourse referents connects with the analysis of pointing in O'Madagain et al. (2019), where it is argued that pointing is usually sight-line pointing, and that such pointing is continuous with touching.





$$(14) \quad wv[p_1 d_2 s_3 s_4 p_5 d_6 e_7 s_8] = \left\{ c \mid \exists x_3 \exists v_2 \exists x_2 \exists x_1 \exists v_1 \left[\begin{array}{ll} c = wv_2 x_3 x_2 x_1 & \wedge \\ \pi(w, v_1) = p_1 & \wedge \\ \bar{\pi}(w, v_1, d_2, x_1) & \wedge \\ \mathbf{name}(w, x_1, \text{"Dick"}) & \wedge \\ \mathbf{dog}(w, x_2) & \wedge \\ \mathbf{have}(w, x_1, x_2) & \wedge \\ \pi(w, v_2) = p_5 & \wedge \\ \bar{\pi}(w, v_2, d_6, x_3) & \wedge \\ x_3 = x_2 & \wedge \\ \mathbf{name}(w, x_3, \text{"Spot"}) & \end{array} \right] \right\}$$

Some comments about the mechanics are in order. In a tuple c of the form $wv_2 x_3 x_2 x_1$, x_3 is a witness for the discourse referent that was introduced last. That discourse referent is introduced by d_2 , and corresponds to the dog in the second picture. x_2 is a witness for the penultimately introduced discourse referent, which is introduced by the phrase [a dog] in s_4 . These discourse referents are distinct, but they are identified by the equation $1 = 2$ in (11b), which equates the ultimately and penultimately introduced discourse referents. This results in $x_3 = x_2$ in the body of (14). v_2 is the viewpoint used for p_5 , and it is also used in selecting values for d_6 , as expressed in the condition $\bar{\pi}(w, v_2, d_6, x_3)$. All the conditions in the body of (14) refer to the same world variable w , when they refer to a world at all. This indicates that the eight parts in (11b) are combining extensionally.

3 The Nominal Depiction Constraint

Look at the matrix of data in (15), where each cell has a picture combined with a nominal English phrase, rather than a sentence. The off-diagonal elements are somehow inconsistent or implausible. For instance, the top right combination with the caption “a castle owned by a duke” is intuitively inconsistent because what is depicted looks like a person, and not at all like a castle. Yet both a duke and a castle are mentioned in the phrase.

(15)

		Drawing of Louis Victor de Rochechouart, Duke of Mortemart, Duke of Vivonne by Antoine Maurin. Wikimedia Commons.
A duke who owns a castle	A castle owned by a duke	
		Castle on a hill. Public Domain Vectors.
A duke who owns a castle	A castle owned by a duke	

The data in (16) are similar. Even though the righthand combination is to some degree pragmatically coherent—the cat basket is empty because the cat that ordinarily occupies it is lost—this combination of a picture and a nominal caption conveys inconsistent information. In contrast, the combination (17), where the caption conveys similar information but is a sentence rather than a nominal, is slightly disjointed but consistent.

(16) a.



A lost cat

b.



A lost cat

Image from page 143 of *The Animals of the World. Brehm's Life of Animals*. 1895. Wikimedia Commons.

Image from page 659 of *Florists' Review*. 1912. Internet Archive Book Images.

(17)



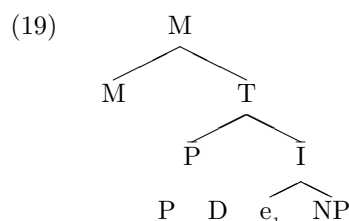
A cat got lost and didn't come home.

These data motivate the *nominal depiction constraint*: roughly, when a picture is accompanied by a nominal, the top-level index in the nominal is co-indexed with a discourse referent pointing into the picture. Or to put it differently, a witness for the top-level index of the nominal is depicted in the picture. For instance, assuming that the semantics of the phrase *a duke who owns a castle* distinguishes a discourse referent for a duke as the top-level index of the nominal, the LF of the top-left combination in (15) should involve a discourse referent pointing into the picture, and this discourse referent should be equated with the top-level index of the caption. This constraint will be imposed syntactically. The syntax of the mixed-medium narratives seen so far can be captured by the context free rules in (18). This creates left-branching trees, consisting of pictures (syntactic category P), sentences (syntactic category S), geometric discourse referents (syntactic category D) and equalities (syntactic category E). M is the syntactic category of cross-medium narratives. A phrase of category S is assumed to be a sentence, as characterized syntactically and semantically by an interpreted grammar of English. So far, this does not introduce any nominal phrases into mixed-medium narratives.

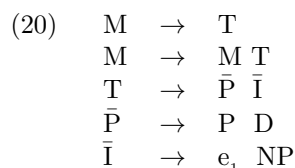
- (18)
- | | | |
|---|---|-----|
| M | → | P |
| M | → | S |
| M | → | M S |
| M | → | M P |
| M | → | M D |
| M | → | M E |

We treat a picture accompanied by a nominal phrase as a special construction that enforces co-indexing. To express this, we hypothesize that the nominal phrases have a predicative

syntactic and semantic type, here assumed to be NP. The tree shape in (19) enforces the required indexing. T is the syntactic category for the construction as a whole. It has two parts. The first part $[\bar{p} \ P \ D]$ is a combination of a picture and a geometric discourse referent. It introduces a picture with a discourse referent pointing into it. Given the recency convention in the dynamic semantics, that discourse referent is accessed with the index 1. The second part $[\bar{p} \ e_i \ NP]$ is a combination of an empty category with index 1 and a nominal predicate with syntactic category NP. It has the effect of applying the nominal predicate to the geometric discourse referent that is introduced in $[\bar{p} \ P \ D]$.



The phrase structure rules covering the construction are in (20). The important point in the analysis of the nominal depiction constraint is that nominal phrases are not introduced freely. Rather they are introduced in a construction that stipulates indexing into a picture.

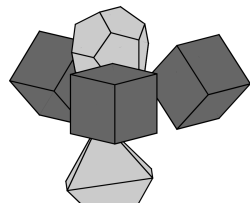


4 Quantification and Definite Reference

In (21a,b), the sentences can be conceived of as observations about the information conveyed by the accompanying picture. (22a,b) are combinations of the same form, but where the sentences give independent information of a kind that can not be conveyed by pictures. In (21a) and (22a), the DPs of the form [every cube] can conceivably be read as quantifying all the cubes in the world. But these DPs are more naturally read to quantify the cubes *that are depicted*. Half of the analysis of this reading is familiar. According to the analysis of Westerståhl (1989) quantificational determiners come with a context variable for a contextually determined domain of quantification. We write this here with a superscripted numerical index. The representation for the sentence in (21a) is then (23a), where the index for the domain of quantification is 1. The value of this index in context should be set in a way that (23a) gets the reading (23b).⁹

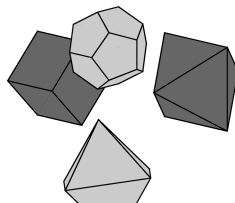
⁹The paraphrase needs to be analyzed too. See the next section.

(21) a.



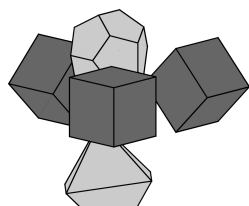
Every cube is dark.

b.



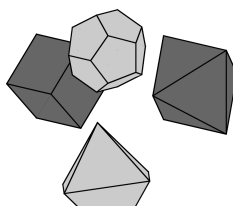
The cube is dark.

(22) a.



Every cube belongs to Jack.

b.



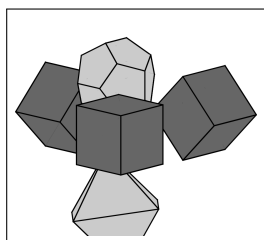
The cube belongs to Jack.

(23) a. Every¹ cube is dark.

b. Every cube that is depicted is dark.

These data lead to the hypothesis that pictures make available or can make available a group discourse referent for the depicted objects.¹⁰ Following the strategy of expressing particular readings syntactically in the discourse representation, we propose an operator *G* that introduces a discourse referent for the group of objects that are depicted in the previous picture. *G* does not involve a point or a bounding box, because it is supposed to introduce a discourse referent for all the depicted objects. It is simply a syntactic constant. (24) is then the discourse representation for the depiction-restricted reading of (21a). The formula is structured linearly, beginning with a picture. Following that the operator *G* introduces a discourse referent for the set of objects depicted in the picture. This discourse referent was introduced last, and so is referenced with the index 1. In the sentence that completes the formula, the index 1 by virtue of its syntactic position contributes the domain of quantification for *every*.

(24)

*G* [every¹ cube is dark]

A semantics for *G* is defined as a quantified version of the semantics of geometric discourse referents. Suppose we are given a picture *p* with unit dimensions, a viewpoint *v*, and a world *w* such that $\pi(w, v) = p$. An object *x* in *w* is depicted in *p* if and only if there is a point *d* in

¹⁰Abusch (2012) also used group discourse referents in analyzing indexing in pictorial narratives.

$[0, 1]^2$ such that $\bar{\pi}(w, v, d, x)$. Therefore we can say that x is a member of the group discourse referent created by G (given w and v) if and only if there is some discourse referent d such that x is a witness for d relative to w and v . This leads to the definition of the semantics of G in (25). Where p_{24} is the picture in (24), (26) is the resulting semantics for (24), where the universal quantification is restricted to depicted objects via the conjunct $y \in X$.

$$(25) \quad wv\mathcal{O}[\![G]\!] = \{c | \exists X. c = wvX\mathcal{O} \wedge X = \{x | \exists d. \bar{\pi}(w, v, d, x)\}\}$$

$$(26) \quad wv[\![p_{24}G[\text{every}^1 \text{ cube is dark}]]\!] = \left\{ c | \exists X \exists v_1 \left[\begin{array}{l} c = wv_1X \\ \pi(w, v_1) = p_{24} \\ X = \{x | \exists d. \bar{\pi}(w, v_1, d, x)\} \\ \forall y [\text{cube}(w, y) \wedge y \in X \rightarrow \text{dark}(w, y)] \end{array} \right] \right\}$$

Examples (21b) and (22b) have sentences with definite descriptions rather universal quantifiers. Here the observation is that the uniqueness presupposition of the definite description is satisfied among objects that are depicted in the picture. For instance in (21b), there is a definite description $[\text{the}_{\text{DP}} \text{ the cube}]$, and in worlds compatible with the picture, there is a unique cube that is depicted in the picture. These examples are analyzed in a parallel way, see (27).

$$(27) \quad \begin{array}{c} \text{[Image of four cubes: one light gray, one dark gray, one medium gray, and one very dark gray]} \end{array} \quad G \text{ [the}^1 \text{ cube is dark]}$$

5 Depiction Sentences

For some of the cross-medium data from Section 3 and Section 4, there are parallel data involving sentences that describe pictures. Recall p_{16a} , the picture of a cat lost in the woods, and p_{16b} , the picture of an empty cat basket. Referring to these pictures, sentence (28a) is true, and sentence (28b) is false, intuitively because it depicts a cat basket rather than a cat.

- (28) a. Picture p_{16a} depicts a cat.
b. Picture p_{16b} depicts a cat.

Sentence (29) is a version of what in Section 4 was cited as a paraphrase of the depiction-restricted reading of a mixed-medium sequence.

- (29) Every cube that is depicted in picture p_{21a} is dark.

These sentences can be used in a discussion among agents who can see the pictures. They can also be used to convey information to an agent who can not see the picture. This makes it implausible that the logical forms of these sentences include particular geometric discourse referents. The reason is that, without access to the picture, a listener can not be expected to accommodate a particular geometric discourse referent. But following the strategy used in the semantics of G , the discourse referent can be quantified in the semantics. This suggests a

semantic paraphrase along the lines of (30) for (28a). It says that there is a discourse referent such that, in every world and viewpoint compatible with the picture, the individual picked out by the discourse referent with respect to the world and viewpoint is a cat.

$$(30) \quad \exists d \forall w \forall v \forall x [\pi(w, v) = p_{16a} \wedge \bar{\pi}(w, v, d, x) \rightarrow \mathbf{cat}(w, x)]$$

This is a formalization of a de dicto reading of the sentence. Although we think this analysis works for pictures of cubes and dodecahedra in a modal space where worlds are occupied only by regular polytopes, cat pictures of the familiar kind do not have information strong enough to entail (30). After all, our own world contains realistic sculptures of cats that are not real cats. Also, depiction sentences have ambiguities along de dicto/de re lines, similar to the ambiguities studied for the verb *paint* in examples like (31) that are studied in Zimmermann (2006). There is much more to say about (28) and (29). Nevertheless, the connection between these examples and the nominal depiction constraint from Section 3 is intriguing, and that connection does fall out of the formalization (30).

$$(31) \quad \text{Edlon painted a bridge.}$$

6 Conclusion

The idea proposed here is to theorize about indexing across media by using a uniform dynamic semantic framework for the media. Indexing is analyzed at the semantic level, where the media are not distinguished. We defined a formal language and a semantic interpretation for it. Particular constructions and constraints were treated in the syntax of the formal language. While it would be possible to do the syntactic part without referring to possible worlds semantics and dynamic semantics, in the research strategy pursued here, the two go hand in hand.

References

- Abusch, Dorit. 2012. Applying discourse semantics and pragmatics to co-reference in picture sequences. In *Proceedings of Sinn und Bedeutung*, vol. 17, .
- Abusch, Dorit. 2014. Temporal succession and aspectual type in visual narrative. *The Art and Craft of Semantics: A Festschrift for Irene Heim* 1. 9–29.
- Abusch, Dorit. to appear. Possible worlds semantics for pictures. In Daniel Gutzman, Lisa Matthewson, Cecile Meier, Hotze Rullmann & Thomas Ede Zimmermann (eds.), *Companion to Semantics*, Wiley. Version of 2015 at hdl.handle.net/1813/44654.
- Abusch, Dorit & Mats Rooth. 2017. The formal semantics of free perception in pictorial narratives. In *Proceedings of the 21st amsterdam colloquium*, .
- Everingham, M., L. Van Gool, C. K. I. Williams, J. Winn & A. Zisserman. 2012. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. pascal-network.org/challenges/VOC/voc2012/workshop.
- Lewis, David. 1975. *Advances of quantification*. In *Formal Semantics of Natural Language*, 178–188. Cambridge University Press.
- Maier, Emar & Sofia Bimpikou. 2019. Shifting perspectives in pictorial narratives. In *Proceedings of Sinn und Bedeutung*, vol. 23 2, 91–106.

- O'Madagain, Cathal, Gregor Kachel & Brent Strickland. 2019. The origin of pointing: Evidence for the touch hypothesis. *Science Advances* 5(7).
- Rooth, Mats. To appear. Adverbs of quantification. In Louise McNally & Zoltán Szabó (eds.), *A Reader's Guide to Classic Papers in Formal Semantics*, Springer.
- Rooth, Mats & Dorit Abusch. 2018. Picture descriptions and centered content. In Rob Truswell, Chris Cummins, Caroline Heycock, Brian Rabern & Hannah Rohde (eds.), *Proceedings of Sinn und Bedeutung*, vol. 21 2, 1051–1064.
- Westerståhl, Dag. 1989. Quantifiers in formal and natural languages. In *Handbook of Philosophical Logic*, 1–131. Springer.
- Zimmermann, Thomas Ede. 2006. Quaint paint. In Hans-Martin Gärtner, Sigrid Beck, Regine Eckardt, Renate Musan & Barbara Stiebels (eds.), *Between 40 and 60 puzzles for Krifka*, Zentrum für Allgemeine Sprachwissenschaft, Typologie und Universalienforschung (ZAS).