Space and the Vision–Language Interface: A Model-Theoretic Approach

Francesco-Alessio Ursini

The relation between spatial vision and spatial language has always been a source of controversy. Three problems can be identified as in need of a solution. A first problem pertains to the nature of the minimal information units that make up spatial vision and language. A second problem pertains to the 'dynamic' aspects of vision and language, or what visual information *to* and similar adpositions correspond to. A third problem pertains to how these different types of information are related one another, and what is the status of this 'interface', especially within a broader theory of language and cognition. The solution proposed here consists in a formal (model-theoretic) treatment of visual and linguistic information, both static and dynamic, that is couched within (a simplified form of) *Discourse Representation Theory*. It is shown that this solution is consistent with general theories of cognition and may shed some (novel) light on the nature of the FLN/FLB distinction.

Keywords: Discourse Representation Theory; faculty of language in the narrow/broad sense (FLN/FLB); interfaces; space; vision

1. Introduction: What We Talk about, When We talk about Space

In this paper, I shall address the problem of the vision–language interface: Informally, what is the exact relation between 'what we see' and 'what we say', (or: "How much space gets into language?"; Bierwisch 1996: 7). This problem can be formulated via the following (and slightly different) global research question:

Q–A: *What is the relation between vision and language?*

I shall suggest that the problem of the vision–language interface and its nature is

I wish to thank Stephen Crain, Max Coltheart, Drew Khlentzos, and Rosalind Thornton for reading and making comments on an initial draft, Qiu-Peng (Autuman) Luo and Miaoling (Anne) Hsieh for thorough feedback on later drafts, the editors of *Biolinguistics* for extensive support during the editorial process, and two anonymous reviewers for their very useful and stimulating comments, which I hope to have addressed in detail. I wish to thank the captains Harlock and Kirk as long-time role models, and my Princess for all her love and support. All remaining errors are of course my own; for those, I apologize to the reader.

ISSN 1450–3417

not much a problem of 'quantity' but 'quality': In order to solve this problem, we need to address not 'how much' information belonging to spatial representations ("what we see") finds its way in language (and *vice versa*), but 'how' this process comes by and how it is possible that visual information can be realized in language in a rather flexible way. I shall argue that in order to understand how sentences such as (1) and (2) can convey non-linguistic spatial information, we need to understand how the relation between "what we see" and "what we say" comes about in the first place.

(1) Mario sits in front of the chimney.

(2) Mario has gone to the rugby match.

This problem can be solved by a *divide et impera* research strategy. I shall first split the problem in three smaller problems (the *divide* part), and solve each of them, integrating these solutions in a 'global' solution (the *impera* part). The three problems that constitute our central problem are the following.

First, we have a *foundational* problem, since previous proposals in the literature make different assumptions on the nature of "what we see'" and "what we say". Some assume that language expresses only shapes of objects (as nouns) and geometrical configurations (as adpositions) (e.g., Landau & Jackendoff 1993); others that we directly express perceptual information "as we see it", without an intermediate level of processing (i.e. language, e.g., Coventry & Garrod 2004). Hence, we don't have a clear (theoretical) picture regarding spatial vision and spatial language, and to what extent they are distinct modules of cognition, let alone a strong, clear theory of their interface.

Second, we have a *descriptive* and *logical* problem, since previous proposals only cover inherently "static" aspects of space, but not "dynamic" aspects. Informally, these theories can account where things *are*, but not where things *are going*. Hence, we do not know what visual information adpositions such as *to* and *from* stand for, nor whether this information should be considered as "spatial'" or not.

Third, we have a *theoretical* and a *philosophica*l problem, since we must define a novel theory that is built upon the solutions to the first and second problem and can explain all the data. Then we must assess the consequences of this theory with respect to a broader theory of vision and language as part of cognition, and their unique aspects — or: What information (and properties thereof) is found in vision but not in language, and *vice versa*.

These three 'smaller' problems can be reformulated as the following research questions:

RQ1: *What do we know so far from the past literature, regarding spatial vision, language and their interface?*

RQ2: *What further bits of spatial knowledge must be included in our models of (spatial) vision and language, and which formal tools used to properly treat these bits?*

RQ3: *What is the nature of the vision–language interface, and which aspects are unique to language?*

Anticipating matters a bit, I shall propose the following answers. First, we know that previous literature tells us that (spatial) vision and language express internal models of objects and their possible spatial relations, and that nouns and adpositions respectively represent objects and possible relations in language. Second, we must include any type of relations in our models of vision and language, insofar as they allow establishing a relation between entities, since the emergent notion of 'space' we will obtain from our discussion is quite an abstract one. Hence, we can use a model-theoretic approach, such as *Discourse Representation Theory* (DRT; Kamp *et al*. 2005), to aptly represent these models. Third, the vision–language interface consists of the conscious processes by which we may match visual representations with linguistic ones and *vice versa*, though some linguistic representations do not represent visual objects, rather 'processes' by which we may reason about these visual objects. Consequently, vision and language can be represented as distinct models sharing the same 'logical structure', which may be connected or 'interfaced' via an opportune set of functions, representing top-down processes by which we may (consciously) evaluate whether what we see accurately describes what we say (or hear), but need not to do so.

This paper is organized as follows. In section 2, I introduce some basic notions and review previous proposals, offering an answer to the first research question. In section 3, I review theories of 'static' and 'dynamic' object recognition, and propose a model-theoretic approach to vision; I then focus on language and offer a DRT treatment of spatial language. In section 4, I integrate the two proposals in a novel theory of the vision–language interface and offer empirical evidence in support of this theory; I then focus on some of the broader consquences of the theory, by sketching an analysis of what properties emerge as unique to language from my theory, thus suggesting a somewhat novel perspective to the nature of the narrow faculty of language (FLN; Hauser *et al*. 2002, Fitch *et al*. 2005). In section 5, I finally offer my conclusions.

2. The Relation between Spatial Vision and Language

In this section I shall outline notions of spatial vision and language (section 2.1) and review previous approaches to their interface, consequently offering the first research answer (section 2.2).

2.1. Basic Notions of Space

Our daily life experiences occur in space and time,[1] as we navigate our environment by analyzing spatial relations between objects. A basic assumption, in cognitive science, is that we do so by processing (mostly) visual information about such objects and their relations as they may evolve over time, e.g., a toy which is

[1] Here and throughout the paper, I shall focus my attention (and use of labels) to 'space', although it would be more accurate to think of our topic as being about spatio-*temporal* vision and language, i.e. how we process location and *change* of location of objects. I hope that the lack of precision will not confuse the reader, and thank an anonymous reader for suggesting this *precís*.

on top of a table, and that we internally represent this information via a corresponding mental 'model' (e.g., Craik 1943, Johnson-Laird 1983, 1992, O'Keefe & Nadel 1978).

Another basic assumption is that, when we share this information with other fellow human beings (i.e. when we speak), we do so by defining a submodel of space in which one object acts as the 'center' of the system, as in (3):

(3) The toy is on top of the table.

With a sentence such as (3), we convey a state of affairs in which, informally, we take the table as the origin of the reference system, take one portion of the table (its top) and assert for the toy to be more or less located in this 'area' (Talmy 1978, 2000). Our cognition of space is thus (mostly) based on the information processed and exchanged between our vision[2] module ("what we see") and our language module ("what we say"). It is also based on an emerging type of information, the structural relations that may be defined between these two modules, our ability to integrate together visual and linguistic units ("what we see and what we say") into coherent representations, over time.

The exact nature of these types of information, however, is a matter of controversy. Some say that spatial vision amounts to information about objects, their parts and shape, and the geometrical relations between these objects as when an object is on top of another (e.g., Landau & Jackendoff 1993, O'Keefe 2003). Another series of proposals offers evidence that other aspects, such as mechanical interactions (a table supporting a toy) and more abstract properties play a crucial role in how we mentally represent space (Coventry & Garrod 2004 and references therein).

We can thus observe that there is a certain tension between 'narrower', or purely geometrical, approaches and 'broader' approaches to both vision and language; as a consequence, there is also a certain tension between theories that consider spatial vision 'richer' than spatial language (e.g., Landau & Jackendoff 1993), and theories that do not assume such difference, often by simply collapsing these two modules into 'cognition' (e.g., Coventry & Garrod 2004). We thus do not have a clear picture of what information is spatial language, and what is spatial vision.

The problem of the exact type of spatial information, however, takes an even more complex nature when we look at another way in which we process spatial information, which can be loosely labeled as 'change'. Take a sentence such as (4):

(4) Mario is going to the rugby stadium.

Intuitively, this sentence describes a state of affairs in which the locatum(s) changes position over a certain amount of time of which we are aware. Mario can

[2] The notion of spatial vision and cognition are somewhat interchangeable for most authors. In this paper I shall use the term 'spatial vision' and 'spatial language' to avoid this confusion. Thanks to an anonymous reviewer for pointing me to this issue.

start at some unspecified starting point, move for a while, and then stop once he's at his planned destination (the rugby stadium). While there are theories of 'dynamic' vision, or how we keep track of objects changing position, as well as theories of 'dynamic' language and more specifically adpositions such as *to*, no one has attempted to integrate these theories into a broader theory of spatial vision and language, let alone in a theory of the vision–language interface.

Another challenge comes from purely linguistic facts, and what kind of information is in a sense 'unique' to a linguistic level of representation. Take a sentence such as (5):

(5) Every boy is going to a rugby field.

In this case, we can have a certain number of boys involved in the corresponding state of affairs, and each of them is described as moving in direction of a rugby field. Yet, if there are several fields at which the children can arrive (Paul goes to Manly's Oval, Joe to Randwick Field, etc.), the sentence may describe slightly different states of affairs, since they informally describe a 'collection' of more specific relations, and what they have in common. As these facts show, we need to take a broader and more flexible perspective in order to address the issue of the vision–language interface than the one usually assumed in the literature, as well as assessing in detail what elements of previous proposals we can maintain in our novel approach. Hence, I am also suggesting that the solution to this problem will offer us a quite different, but hopefully correct, answer to the 'problem of space'. Before offering this answer, however, I shall review the previous literature.

2.2. Previous Literature

Previous proposals on the vision–language interface can be divided into a 'narrower', geometric approach (or: "spatial language expresses geometric relations") and 'broader', 'functional' approach (or: "spatial language also expresses extra-geometrical relations"). One well-known and influential example of the geometric approach is Landau & Jackendoff (1993, henceforth L&J), while a well-known and influential functional approach is the *Functional Geometric Framework* (FGF; Coventry & Garrod 2004). I will offer a review of both, highlighting their features and shortcomings, with respect to the topic of this chapter, starting from L&J's proposal.

L&J offer evidence that, at a visual level, objects and their relations are captured using "spatial representations", chiefly expressed by adpositions. Size, orientation, curvature and other physical properties all conspire for an object to be recognized as more than a sum of its parts: a 'whole' entity, or what the object is. Whole objects or '*what*s' can also be related one to another: If we have two objects, one will be conceived as a landmark object (or *ground*), while the other will be the 'located' entity (or *figure*; Talmy 1978, 2000).

They also argue that the rich and variegated layers of visual-cognitive information are processed and then clustered together and associated with 'conceptual labels' (or just 'concepts') and hierarchically organized within the

Conceptual System (CS, Jackendoff 1983, 1990, 1991, 2002), the interface between non-linguistic modules and (linguistic) domain of semantics. This proposal and further extensions assumes that *nouns* are the main category representing objects in language, whereas *adpositions* represent spatial representations/relations (e.g., van der Zee 2000). In line with other literature, L&J propose that spatial expressions mostly involve 'count' nouns, which can be seen as labels for objects with a given 'shape' (e.g., 'cylinder' or the fictional 'dax': Carey 1992, 1994, 2001, Soja *et al.* 1992, Bloom 2000, Carey & Xu 2001). Adpositions, on the other hand, are argued to express core geometrical properties such as overlapping, distance and orientation (e.g., *in*, *in front of*; Landau & Stecker 1990, Landau *et al.* 1992).

Recent inter-disciplinary research has shown that the picture is somewhat more complex. A rich body of evidence has been accumulated suggesting that adpositions can also convey information which is not necessarily geometric in nature. Look at the examples:

(6) The book is on the table.

(7) Mario is beside the table.

(8) #The table is beside Mario.

(9) Mario is taking the moka machine to the kitchen.

If a book is "on" the table (as conveyed by (6)), the table will also act as a mechanical support to the book, that is, it will prevent the book from falling. We can say that Mario is "beside" the table (as in (7)), but saying that the table is beside Mario will be pragmatically odd (as in (8)):[3] Figures tend to be animate entities (or at least conceived as such), whereas grounds tend to be inanimate entities.

These mechanical properties can also be seen as extra-linguistic or 'spatial' properties associated to nouns. Informally, if a count noun such as *book* is associated to an object with definite shape, it can (and should) be involved in causal physic relations (e.g., support, or containment); cf. Kim & Spelke (1992, 1999), Spelke & van der Walle (1993), Spelke *et al.* (1994), van der Walle & Spelke (1996), Spelke & Hespos (2001), Smith *et al.* (2002), Shutts & Spelke (2004).

Dynamic contexts offer similar evidence for the relevance of extra-geometric information to be relevant. For instance, in a scenario corresponding to (9), we will understand that the Moka machine[4] brought to the kitchen by Mario will reach the kitchen because of Mario's action (Ullman 1979, 1996, von Hofsten *et al.* 1998, 2000, Scholl 2001, 2007). We will also take for granted that the machine's handle and beak will reach the kitchen as well, as parts of the machine, unless some problem arises in the meanwhile. If Mario trips and the Moka machine falls mid-way to the kitchen, breaking in many pieces, we may not be able to recognize the Moka machine as such (Keil 1989, Smith *et al.* 1996, Landau *et al.* 1998). Spatial relations, and thus adpositions that express these relations, can implicitly capture the (potential) causal relations or affordances between dif-

[3] Examples (7) and (8) and related discussion are based on an issue correctly pointed out by an anonymous reviewer, whom I thank.

[4] The traditional Italian machine for espresso coffee.

ferent objects (e.g., Landau 1994, 2002. Munnich & Landau 2003).

For these reasons, Coventry & Garrod (2004) propose their FGF framework, according to which mechanical, geometrical and affordance-oriented properties form the *mental model* or *schema* (in the sense of Johnson-Laird 1983) of adpositions that we store in *long-term memory*. This model can be seen as the 'complete' representation of an adposition's meaning, which can then only partially correspond to its actual instantiation in an extra-linguistic context (see also Herskovits 1986).

According to this theory, speakers can then judge a sentence including a spatial adposition more or less appropriate or felicitous, depending on whether the adposition's content is fully or partially instantiated in an extra linguistic scenario (e.g., van der Zee & Slack 2003, Coventry & Garrod 2004, 2005, Carlson & van der Zee 2005, Coventry *et al.* 2009, Mix *et al.* 2010). Two examples are the following:

(10) The painting is on the wall.

(11) The painting is in the wall.

A sentence such as (10) can be considered more appropriate than (11) when used in an extra-linguistic context in which a certain painting is just hanging on the wall, but less appropriate when the painting is literally encased in the wall's structure.

Other theories take a perspective which is either close to L&J or FGF. *Vector Grammar Theory* (O'Keefe 1996, 2003) treats English adpositions as conveying information about vector fields, the graded sequence of vectors representing the minimal 'path' from ground to figure, and thus conveying purely geometric information. Another theory which is based on similar assumptions is the *Attentional Vector Sum* model (AVS; Regier & Carlson 2001, Regier & Zheng 2003, Carlson *et al.* 2003, 2006, Regier *et al.* 2005). In this theory, 'vectors' represent features of objects that can attract the speaker's attention once he interprets a spatial sentence, and can thus include mechanical and functional aspects as well as environmental ('reference frames') information.

These theories thus predict that a sentence such as (12),

(12) The lamp is above the chair.

is interpreted as a 'set of instructions' that informs us about where to look at, in a visual scenario, but they differ with respect to these instructions being purely geometrical or not. Furthermore, AVS predicts that *above* will be considered more appropriate if used in an extra-linguistic context in which the lamp is above the chair also with respect to three possible systems of orientation or reference frames, for example, if the lamp is above the chair with respect to some environmental landmark such as the floor (*absolute* reference frame), with respect to the chair's top side (*intrinsic* reference frame), and with respect to the speaker's orientation (*relative* reference frame); see e.g. Carlson-Radvansky & Irwin (1994), Carlson (1999).

Although the insights from these theories are quite enlightening and consistent with various approaches to vision, their approach to language is inherently a 'blurry' one, as each of these theories says virtually nothing about the specific contribution of nouns and adpositions. Since these theories tend to reduce language to general cognition, this is not surprising. Aside from this problem, no theory really attempts to analyze 'dynamic' spatial expressions. The same holds for L&J and FGF: Examples such as (4) and adpositions such as *to* are still a mystery, with respect to the vision–language interface. Nevertheless, both sides of the debate offer at least two important points regarding the nature of spatial vision and spatial language.

These aspects form the answer I shall propose to the first research question:

A–1: *Previous literature offers a clear mapping between vision and language (L&J), and evidence that spatial vision and language express possible relations between entities (FGF).*

Because of these previous proposals I shall assume, based on the literature on the topic, that spatial vision and spatial language are not just about geometrical relations, and thus suggest that both modules can express the same 'amount' of spatial information, although in (quite) different formats. I shall also assume that there is one precise, although flexible, correspondence between units of vision and units of language. Visual objects find their way in language as nouns, and spatial relations as adpositions, at least for English cases I shall discuss here.[5] In the next section, I shall offer a justification to these assumptions and propose a richer theory of spatial vision and language.

3. The Nature of Spatial Vision and Language, and a Formal Analysis

In this section I shall offer an analysis of 'static' and 'dynamic' vision (sections 3.1 and 3.3), and a logic of vision of these theories (sections 3.2 and 3.4); I shall then analyze (specific aspects of) spatial language via DRT (section 3.5).

3.1. Classical and Modern Varieties of Object Recognition

In highly schematic terms, we can say that spatial information is processed via visual perception, for most human beings. Light 'bounces' off an object and the surviving wave-length is processed by the eyes. This information is then transmitted to the optic nerve, to be further processed in various parts of the brain, like the primary and secondary visual cortex. Once the perceptual inputs are processed, their corresponding (internal) representations become the basic chunks or atoms of information processed by higher cognitive functions, such as vision and

[5] A specific language may lack a term for a certain visual object, so the correspondence between visual objects and nouns on the one hand, and spatial relations and adpositions on the other hand, may be subject to subtle cross-linguistic variation. Informally, if a language has a term for a certain visual object, this term will be a noun, syntax-wise: The same holds for spatial relations. I thank an anonymous reviewer for bringing my attention to this point.

memory.

One of the earliest schools of research that attempted to investigate the nature and properties of these units of information was the *Gestalt* school of psychology. This school assumed that our unconscious processes of visual recognition allow us to individuate objects from the background via the following four principles: *invariance* ('sameness' of an object), *emergence* (parts making up a whole), *reification* (interpolation of extra information), and *multi-stability* (multiple 'good' images of an object). These principles converge into underlying principle of *Prägnanz* or *conciseness*, our ability to form discrete visual units from different, and perhaps contradictory, 'streams' of perceptual information. This process may not necessarily be 'veridical' in nature: If we look at a car in motion and we do not notice its radio antenna, we may consider the two objects as one, as long as there is no visual cue that they are indeed distinct objects (e.g., the antenna breaks and flies away).

The Gestalt school's thrust in the study of invariant properties lost momentum after the end of World War II, until Gibson (1966) re-introduced the study of vision as a process of 'information-processing' (and integration), which sparked the interest of various researchers,[6] including David Marr and his model of vision which had an ever-lasting influence in vision sciences and in some linguistic literature (e.g. van der Does & van Lambalgen 2000).

Marr's initial research started from the physiological bases of vision (collected in Vaina 1990). His interest slowly shifted from the neurological and perceptual facts to cognitive aspects of visual processes, which culminated in Marr (1982). The core assumption in Marr's theory is that vision can be best understood and represented as a computational, algebraic model of information processing. It is a bottom-up and cognitively impenetrable process, since it is mostly realized without the intervention of conscious effort.

Marr proposed that any model, and thus any mental process or structure it represents, should be defined at three levels of understanding: *computational* ("why" of a model), *algorithmic* (the "how" of a model), and *implementational* (the "what" of a model). Marr proposed that our vision developed with a perhaps very abstract computational nature, that of 'grouping' any type of visual information (geometric and not) into implementable units, which can be retrieved and stored in memory. Regardless of its purposes, Marr proposed that the computational system of human vision is assumed to have three intermediate levels of representation, or 'sketches'.

At the *Primal Sketch* level, boundaries ('zero crossings') and edges are computed, so that the continuous stream of perception is partitioned into discrete units of attention, or 'receptive fields'. Photo-receptive cells detect the change of light in the receptive fields, and split it in two parts: an 'on-center' and an 'off-center'. In 'on-center' cells, the cell will fire when the center is exposed to light, and will not fire when the surround is so exposed. In 'off-center' cells, the opposite happens. When both types of cells fire at the same time, they are able to

[6] J.J. Gibson would come to reject his stance in favor of an 'ecological' or 'externalist' approach, in Gibson (1979). More information about perceptual and historical aspects can be found in Scholl (2001, 2007), Bruce *et al.* (2004), and Farah (2004), *inter alia*.

represent an entity like an edge, its adjacent 'empty' space and the boundary between the two partitions. The change of polarity between these two partitions is defined as a *zero-crossing*. A zero-crossing represents change in terms of opposed polarities: if an edge is marked as +1 in value, then the adjacent 'empty' part will have value –1, and a border will be represented as 0, or as a 'boundary'.

At the *2½-D sketch* level, these elements are integrated in the computation of surfaces and their distance from the observer. For instance, a triangle represents three lines whose edges coincide in a certain order, forming a connected contour, the triangle itself. Other information, such as depth or orientation, is computed via the integration of information about, respectively, the distance of the single surfaces from the observer (hence, an *egocentric* perspective), and integrated in a mean value, the normal 'vector' from those surfaces. Missing information can here be interpolated: If part of the triangle's side is occluded, we may just 'infer' it from the orientation of the visible sides.

At the *3-D model* level, the recognized parts and portions are integrated into one coherent whole. At this level, vision becomes an object-centered (or *allocentric*) process, which allows for shape recognition to be viewpoint-invariant. The computation of a full 3-D model (object recognition) is crucially based on how the computation evolves from the 2½-D sketch to its final level. If the various 2½-D sketches can be integrated into a coherent unit, and this computed unit matches with a corresponding unit in memory, then the process of 'object' recognition is successful (see also Marr & Nishihara 1978).

Marr's model, given its algebraic nature, can be informally stated as a model in which basic information units or indexes can represent single parts of an object: *a* and *b* can stand for head and torso of a human figure, represented as the index *c*. If the unification or merging[7] of the two more 'basic' information units *a* and *b* into a single unit is identified with a whole, then object recognition occurs. Simply put, from head and torso (and other parts) we obtain a human figure, a process that can be represented as $(a{+}b){=}c$, *c* standing for the human figure index.

This quite informal exposition should already made clear that two basic principles can be identified as being part of spatial vision. One is the need to 'chunk' the perceptual stream into discrete, computational units; and the other possibility to 'merge' and identify these units in a rather abstract way, which allows us to establish *part-of* relations, according to Marr, among different information units.

After Marr's seminal work, theories of object recognition roughly distributed between a more representational and a more derivational stance. While representational theories stress relations between different objects and parts (or, rather, representations thereof), derivational theories stress the processes by which these representations come into being. I will start from the representational stance, introducing *Recognition By Components* theory (henceforth RBC; Biederman 1987, Hummel & Biederman 1992), probably the most influential theory for the representational stance.

[7] Here I use the term 'merge' in a pre-theoretic way, but I will offer a more precise definition in section 3.3.

RBC offers an approach which is substantially similar to Marr's original proposal, although it is postulated that object recognition occurs via 7 sketches of representation, rather than 3. One important difference is that, after the first two sketches are computed, each (part of an) object is conceptualized as a *geon* (*generalized ion*; Biederman 1987), a primitive shape or visual 'ur-element'.[8] The combination of various geons allows to define complex forms: For instance, an ice-cream can be idealized as a semi-sphere connected to a cone, consequently capturing complex relations between the parts they represent. Whenever an object is successfully recognized, it can be and stored in memory as a distinct entity (Hummel & Stankiewicz 1996, 1998, Stankiewicz & Hummel 1996).

An important aspect of RBC is that it addresses how different information units are combined together over the time of a computation, a phenomenon defined as dynamic binding. Informally, if we recognize a sphere shape *a* and a cone shape *b* at a(n interval) time *t* in the computation, their integration as integrated units *a+b* will occur at a time *t+1*. In this perspective, object recognition can be seen as a dynamic process of binding different units of information together, so that 'new' objects emerge from this process: By dynamically binding edges and lines together in a coherent representation we have surfaces, and by dynamically binding surfaces together we have three-dimensional objects, at an interval *t+n*.

An alternative view to this representational approach may be exemplified by the derivational model *H-MAX* (short for 'Hierarchical MAXimization' of input) of Tomaso Poggio and associates (Poggio & Edelman 1990, Riesenhuber & Poggio 1999a, 1999b, 2000, 2002, Serre *et al*. 2005). In this model, objects can be any parts of which we receive visual input, via their luminosity, and of which we compute possible visual candidates (e.g., different possible representations of the same dog). No intermediate levels of representation are however assumed to exist, since the flow of information is constrained via a pair of simple principles, *SUM* and *MAX*, which are in turn defined over vectors as sequences of minimal parts and boundaries of an object.

An example is the following. Suppose that we look at our pet Fido, starting from his tail. At this initial step, our visual system first computes parts and boundaries, such as the tail's tip, which can be badly lighted or 'stilted', if we are observing it by an odd angle. From this 'vector', we access other possible memorized images of Fido's tail and combine them with other visual features (vectors) we recognize about Fido. In case the image is somehow poor, we may compare it as a 'noisier' version of Fido's tail.

All these vectors are then summed together in the sum vector, the averaged sum of the vectors corresponding to the various visual inputs. If this sum exists, then a 'standard' (or allocentric) view will be defined, which corresponds to the final step of the process of object recognition. In keeping track of these different views, 'feature clusters', edges of a surface or other easily observable points play

[8] Geons are not exactly primitives *per se*, but represent the (finite) set of combinations (36 in total) of 5 binary or multi-valued properties that combine together to define a shape. These five properties are: *curvedness* (if a component is curved or not), *symmetry*, *axis* (specifically, the number of axes), *size*, and *edge type* (if the edges define an abrupt or smooth 'change of direction').

a vital role.

In more formal terms, the *SUM* takes two visual objects and unites them together into a new visual object: If *a* and *b* are Fido's head and torso, then *a+b=c* is Fido's body. The *MAX* operation minimally differs from the *SUM* operation in two subtle ways. First, it may sum together two visual objects and obtain one of the two objects as the result, i.e. *a+b=b*. This is possible when one object 'includes' the other, i.e. when one visual object contains all the features of another object; hence, their union will be the 'strongest' object. Second, it may average visual objects representing the same entity, i.e. it may sum objects which have common features. In formal terms, this can be then represented as *(a+b)+(b+c)=a+b+c*, a novel visual object (the 'average' image) obtained out of previous objects. These processes are dynamic, so if two visual objects are *SUMmed* (*MAXed*) at a time *t*, the result will hold at a time *t+1*.

While these two theories show a substantial convergence in their treatment of object recognition, their assumptions about the nature of 'objects' is quite different. Representational theories consider an 'object' as the end result of a visual computation, while derivational theories consider an 'object' as any unit that is manipulated by a computation. This difference may appear purely theoretic, but it has its own relevance once we take in consideration how this information is mapped onto linguistic units. Consider, for instance, the following examples:

(13) The book is on the tip of the left edge of the blue table.

(14) The book is on the table.

In (13), the spatial relation is defined over a book and a rather specific part of a blue table, the tip of its left edge, whereas such level of detail is left implicit in (14). Note that this relation also informs us that the book is supported by one part of the table (the tip of the left edge), which in turn may be seen as not so ideal for supporting books (tips are intuitively worse 'supports' than centers).

For the time being, though, I shall leave aside adpositions and spatial relations, and concentrate on objects and nouns. In both sentences, any object or part thereof ('edge', 'tip') finds its linguistic realization as a noun: If there is a difference between different layers of visual representation, this difference disappears at a linguistic level, since both visual objects are represented in language as nouns. Consequently, a theory of object recognition that makes no difference between parts and whole objects, such as *H-MAX*, offers an easy counterpart to these simple linguistic facts, while other theories are less suitable for my goal of offering a theory of the vision–language interface. I shall base my formal proposal on vision by offering a logical treatment of *H-MAX*, in the next section.

3.2. A Logic of Vision, Part I: Static Vision

The core aspects shared by the models of static vision (object recognition) we have seen in the previous section are the following. First, vision involves the explicit, internal representation of perceptual stimuli in terms of discrete information units, or visual objects (of any size and shape, so to speak). Second, these units are combined together via one underlying principle, which we can temporarily label as 'sum'. Third, the result of this process defines more complex objects, but also relations between these objects, which can be seen as instances of the *part-of* relation. These three aspects can be easily represented in one (preliminary) unified logic of vision, which I shall define as follows, and which I shall expand in more detail in section 3.4.

First, I shall assume that vision includes a set of *visual objects*, the (countably infinite) set $V=\{a,b,c,...,z\}$. Each of these objects represents a minimal information unit, an output which is activated (instantiated) when some perceptual input exceeds a threshold level. Hence, each information unit in a computation represents an instance of *transduction*, since it represents the (automatic) conversion from one type of (input) information to another type of (output) information (Pylyshyn 1984, Reiss 2007). I shall assume that each object can be represented as a singleton set, via 'Quine's innovation': Hence, a is shorthand for $\{a\}$; consequently, our operations will be defined over sets (cf. Schwarzschild 1996: appendix).

Second, I shall assume that one syntactic operation can be defined over these units, the sum operation '+', an operation that I will call *merge*. An example of merge is $a+b=c$, which reads: "c is the merge of a and b". It is a *binary* operation which is also *associative*, *commutative*, and *idempotent*. Associativity means that the following holds: $a+(b+c)=(a+b)+c$. In words, and using again the example of Fido, Fido's head with Fido's body (torso and legs) correspond to the same object as Fido's upper body and legs: Fido. Commutativity means that the following holds: $a+b=b+a$. In words, Fido's head and body form Fido, much like Fido's body and head. Idempotence means that the following holds: $b+b=b$. Fido's head and Fido's head give us Fido's head, that is, we can repeat information. Since our objects are singleton sets, this operation is basically equivalent to *set union*. The intuition behind the *merge* operation is that it takes two 'old' distinct objects and creates a 'new' object as a result, in a sense distinct from the basic sum of original parts. For instance, our Fido can be conceived as the new visual object that is obtained when the visual objects corresponding to Fido's body and Fido's head are merged together into an integrated representation, Fido as a 'whole' entity.

Third, I shall assume that one semantic relation can be defined between objects, the *part-of* relation, represented as '≤'. An example of the *part-of* relation is $a \leq b$, which reads: "a is part of b". Since I am using Quine's innovation, the *part-of* relation is roughly equivalent to set membership.[9] This relation is also binary, and it is *reflexive*, *transitive* and *antisymmetric*. It is reflexive, since the following holds: $a \leq a$. It is transitive, because the following holds: if $a \leq b$ and $b \leq c$, then $\mathrm{a} \leq c$. It is antisymmetric, because the following holds: if $a \leq b$ and $b \leq a$, then $a=b$. In words, each part of Fido's is part of itself (reflexivity); if Fido's leg is part of Fido's body and Fido's body is part of Fido, then Fido's leg is part of Fido (transitivity); if Fido's body parts are part of Fido, and Fido consists of Fido's body parts, then

[9] The subtle but very important differences between the notion of 'set membership' and the *part-of* relation are not important for our discussion. However, the interested reader is deferred to e.g. Link (1983, 1998), Landman (1991: chap. 1), Schwarzschild (1996: chap. 1) for discussion.

they are recognized as the same entity (antisymmetry). The intuition behind the *part-of* relation is that it establishes a relation between 'old' objects and a 'new' object as a result of the *merge* operation. For instance, if Fido is the result of merging Fido's legs and Fido's body into a 'new' object, then Fido's legs will be part of Fido. If we recognize Fido, then we will also recognize Fido's legs as well as other parts that make up Fido, as a consequence of the relation between parts and whole.

The resulting model of (object) vision emerging from these basic definitions is the triple $S=<V,+,\leq>$, a simplified variant of a structure known as join lattice, a type of full Boolean algebra (e.g. Keenan & Faltz 1985: chap. 1, Landman 1991: chap. 2, Grätzer 1978: chap. 1–2). A join lattice can be seen as a set with at least one binary operation of composition and one relation defined over its elements, which also has the following property: if $a\leq b$, then $a\cap b=a$ and $a\cup b=b$. In words, if a is part of b, then the intersection of a and b is a, while the union of a and b is b. Informally, if the merge of two objects creates a novel object, the part of relation establishes that this novel object includes the old objects as its (proper) parts. Because of these properties, this type of Boolean algebra is a complete structure, i.e. it will have one maximal object including every other object (i.e. V) and one minimal object which is included in every other object, which we will call '*0*', and which represents any instance in which we 'fail' to recognize objects.[10]

Since we mostly operate on individuals, i.e. singleton sets via *merge* and the *part-of* relation, the logic of vision I define here is substantially a first order logic. Since this logic allows us to define an algebraic model of objects and their interpretation and relations, it is a model-theoretic approach to vision. Anticipating matters a bit, the discussion of the vision–language interface will coincide with the discussion on how this model and the model defined by language are related.

These logical/algebraic properties represent the following facts: The visual 'integration' of Fido's leg and Fido gives us Fido, i.e. Fido's leg is 'recognized' as part of Fido's whole image (union). If from Fido's whole image we focus on Fido's leg, then the other parts will be ignored (intersection). This latter interpretation of 'attention as intersection' can be found in RBC and Ullman (1996), and is based on one simple intuition: If *merge* represents object recognition (the union of different visual) inputs, then its complementary operation represents the process by which we focus on a single visual object out of an array of objects, i.e. attention. Furthermore, the sum of objects forms the full 'library' of our model of vision (the maximal object V), and there can be cases in which we cannot recognize any object whatsoever, for instance when we fail to focus our attention on something (the empty object).

This brief and semi-formal *excursus* suffices for our discussion of object recognition. The important aspect is that we can now define a tight relation between the syntax and semantics of our logic of vision: For each instance of the

[10] Note that the operation *MAX* can be now reconstructed as a special instance of *SUM* (i.e. our merge). I shall leave to the reader the simple proof of this fact. Also, note that given our definition of the sum operation, visual objects can be either *atomic*, i.e. they only include themselves as proper parts (e.g., *{a}*), or non-atomic, they may have other objects as their proper parts (*plural/sum* objects: e.g., *{a,b}*, including *{a}* as its part; see e.g. Link 1983, 1998 and Schwarzschild 1996). The import of this subtle distinction is not crucial, in this paper.

merge operation, the result will define another visual object and a *part-of* relation between this object and its constituent parts. Informally, we are able to recognize the legs of Fido as Fido's, because we first integrate Fido's legs with other Fido's body parts into Fido's whole image, and then retrieve this relation between legs and Fido.

The merging of visual objects does not occur in a temporal void, as we have seen, but is dynamically realized over discrete intervals of time. In RBC, this is represented via *dynamic binding*, i.e. the explicit representation of derivations as they occur over time. Before defining dynamic binding, I shall define the structure of the Index Set that represents intervals of time. This structure is the duple $I=<t,+>$, a set of intervals of time with an operation of *addition*. Although I represent this operation via '+', it is a slightly different operation than *merge*, since it is only associative but not commutative nor idempotent. Intuitively, from a starting interval t we can 'move forward' to other intervals, e.g., $t+1$, $t+2$ and so on, via the simple iteration of this 'asymmetric' *merge*.

The corresponding type of structure is a simpler algebra, a *total order*, i.e. a structure in which each element is a distinct object. Intuitively, this structure represents the directed flow of the logical processes underpinning visual computations, the 'arrow of time' that tells us how visual objects are integrated together, but which cannot 'remember' any relations between the objects manipulated in these operations.

The explicit integration of this structure with vision is the duple $S_d=<I,S>$, the 'dynamic' logic of vision and object recognition. Its dynamic nature stems from the ability to represent visual computations as they occur over derivational times, in a simple format similar to standard proof-theoretic (i.e. syntactic) component of various logical systems (see e.g. Landman 1991 for discussion). One example is the following:

(15) $t.$ a (visual object instantiation, e.g. Fido's head)
$t+1.$ b (visual object instantiation, e.g. Fido's body)
$t+2.$ $a+b$ (*merge* introduction)
$t+3.$ $(a+b)=c$ (Fido as 'sum' of Fido's parts)
$t+4.$ $a\leq c$ (*part-of* introduction, Fido's head as part of Fido)

This derivation roughly captures how the process of recognizing Fido may occur a dynamic (and bottom-up) way, modeling the actual processes described in the reviewed theories. The various objects are first recognized ('instantiated' in the derivational space) one by one and then merged via the introduction of this operation. Once this process is over, we can also access the relation between Fido's head and Fido's whole image, since we can establish that one is part of another.

This simple example of a derivation in our logic of vision may not capture all the aspects involved in visual computations and, to an extent, it is quite idealized: For instance, an individual may consciously assume (and thus exert a *top-down* choice) that he is seeing Fido's body, since he can partially recognize it as a visual entity connected to Fido's head. In this and other examples, I shall leave these matters aside, as they are not crucial, for our discussion. This example, however, introduces one important advantage of my theory over the theories I reviewed so far: it makes fully explicit the structural relations between the various components of the object recognition process, including its unfolding over time. This logic of vision is still a preliminary proposal, since for one thing, it does not allow us to make a distinction between objects (individual constants such as e.g., a) and the properties they instantiate (e.g., constant functions such as **dog′**). It also cannot represent spatial representations, and thus the visual content of adpositions, but this is a void that will be filled in section 3.4, along with a theory of visual properties. However, it already allows us to give a compact definition on how we see things in the world, at least with respect to static objects.

Now we can explicitly represent (visual) objects in a very preliminary *logical space*, and we can also define how these objects are mapped onto their corresponding linguistic labels, nouns. I shall assume, differently from previous proposals such as L&J, that this mapping is an *isomorphism*, a one-to-one correspondence between objects of different types (i.e. visual objects to noun labels). The reasons for this assumption are the following. The discussion of examples (13) and (14), and the intuition that each visual object may (potentially) have a corresponding 'noun' label, has one important theoretical consequence. If we define a function mapping visual objects to nouns, then this function will be *injective*, it will find at least a label $\boldsymbol{n'}$ for each visual object v: A noun like *table*, for instance, stands for the corresponding visual object, a table. Furthermore, it is possible that several visual objects can correspond to one linguistic label: A noun such as 'table' also stands for the sum of legs, surface, edges, and other visual objects making up a table. Hence, this mapping function will be *surjective* as well.

A function which is injective and surjective is a *bijective* function, hence a function that defines an isomorphism. More formally, for each visual object v, for each noun label $\boldsymbol{n'}$, there will be a function f such that : $f(v)=\boldsymbol{n'}$**.** Since this function is surjective, the following holds: given $a+b+c=v$ then $f(a+b+c)=\boldsymbol{n'}$. In words, we have the 'lexical' identity $\boldsymbol{edge'}+\boldsymbol{legs'}=\boldsymbol{table'}$, which can be also indirectly represented as $f(a+b)=f(a)+f(b)$, with $f(a)=\boldsymbol{edge'}$, $f(b)=\boldsymbol{legs'}$ and $f(a+b)=\boldsymbol{table'}$. Furthermore, this isomorphism preserves relations, so if one object is part of another, one corresponding noun will be lexically related to another. We have $f(a)\leq f(b)$, which in words says that *edge* is (lexically) related to *table*.

This isomorphism can be interpreted as follows. Our logic of vision is a partial, yet very fine-grained model of object recognition, with a simple yet rich *hierarchical structure*, defined by the *part-of* relations that can be established between the objects in this domain. The function f tells us that such structure can also be connected with other structures, provided that they are governed by the same (logical) principles. Informally, it allows us to potentially define a correspondence between nouns in language and visual objects in vision, on a one-to-one basis. Although a language may lack a specific lexical item for each visual object, it is at least possible to define such a tight correspondence between nouns on the one hand, and visual objects on the other hand.

This function can be thus thought as representing a top-down, conscious (and optional) process, which occurs when we consciously match visual information against linguistic information. It allows to define a correspondence between simple and complex visual objects and the nouns that represent these objects at a linguistic level, e.g., to establish that a noun such as table can indeed refer to[11] a visual object we may observe, and which is made of four legs, a surface and other relevant parts. With this notion in place, then, we have introduced enough 'machinery' to handle the static side of vision and its logic; we need to focus on the neglected dynamic side, and propose a full logic of vision, by which we can also analyze spatial representations/relations. I shall do so in the next two sections.

3.3. Theories of Dynamic Vision

In the discussion in the two previous sections, I have introduced a view of spatial vision in which the ability to explicitly represent objects and their relations plays a crucial part in 'static' scenarios, i.e. cases in which we 'find' objects which are not changing position over time. One aspect missing from this discussion is how we establish relations between objects, especially when they change position over time — how dynamic spatial vision comes about.

A preliminary step to answer these questions is to define how we can keep track of objects over time. For this purpose, I shall review a theory about dynamic object tracking: *Multiple Object Tracking* (MOT), introduced in Pylyshyn (1989) and developed in a number of successive works (e.g. Pylyshyn 1994, 2001, 2003, 2004, 2006, and Pylyshyn & Annan 2006; see Kahneman *et al.* 1992 for the roughly equivalent *Object File Theory*).

MOT offers a theory about object recognition in dynamic scenarios by analyzing how we are able to individuate and form mental representations of objects in the way they instantiate some properties (e.g., being yellow in color), and by how we maintain or change these representations over time and the unfolding of events. MOT is probably best presented via a preliminary example. Imagine that we look at the panorama: We detect trees, clouds, buildings, and so on. If we focus our attention on a flying black swan, we can do so because we are first able to detect a mysterious object (call it '*x*'), which instantiates the properties "swan", "black", and "flying", among others.

With some imagination, we can assume that "swan" is the primitive and most basic property which allows us to recognize the mysterious entity as such, the equivalent of an imaginary finger stretching from our eyes to the object itself. Such a finger allows us to define the mysterious object in terms of what property it instantiates, and it is thus defined as *Finger of INSTantiation*, or *FINST*. The very act of this process is usually defined as *FINSTing* in the literature and, since it can be defined for any entity that can be so individuated, it makes no distinction between types of objects: Everything which can be *FINSTed* is an object, simply enough.

It is useful to illustrate MOT's notation for the basic process of FINSTing, as well as the addition of further features. I will follow Pylyshyn's (1989) notation, for ease of exposition. Aside from the basic process of FINSTing, we can imagine

[11] The notion of 'reference' I use here is not equivalent to the one commonly employed. A standard assumption is that reference is the relation between a term and the 'real world' object that corresponds to a given term. Here and for the rest of the paper, I shall assume that linguistic terms can refer to extra-linguistic but internal information, such as visual objects.

a situation in which the black swan is flying above a cloud. The process of FINSTization is illustrated in (16), while (17) illustrates the more complex 'above' case:

(16) a. *FINST[x],[swan]=(x:swan)*
b. *FINST[x:swan],[x:black]=(x:swan,x:black)*

(17) *ABOVE(x:swan, x:black, x: f lying, y:cloud)*

In (16a), a basic property like "swan" is mapped onto a visual object, acting as the FINST that tracks the visual object. In (16b), the combination of two properties acting as FINSTs creates a new, more complex FINST, which identifies the visual object *x* as a black swan. In the case of (17), we can observe that such 'internal fingers' can also define relations between simpler 'fingers', hence expressing a relation between different instances of the same underlying process.

This relation is, in turn, a description or property of an event of motion, in which the swan is the moving figure, while the cloud is the contingent ground. Further information can be stacked up via dynamic binding: Informally, each individuating property for *x* can be in a temporally incremental fashion (e.g., "black" at time *t*, "flying" at time *t+1*), which in turn is realized via the iterated application of the FINST operation.

One problem emerging from the presentation of MOT is that this theory cannot easily be used to analyze how the temporal relations between properties can be defined and represented in their own right. While "black" may be instantiated after "swan", we cannot explicitly represent that the corresponding 'fingers' can be taken as entities in their own right, the events during which these properties are instantiated and combined together, or defined in terms of their order of occurrence.

One theory that aims to fill this conceptual void is *Event Segmentation Theory* (henceforth: EST), a theory of events and psychological events first outlined in Zacks & Tversky (2001) and Zacks *et al.* (2001). In this theory, an original philosophical intuition by Quine (1960) and further developed in Davidson (1967) acts as the basic insight and ontological assumption: that our understanding of the world includes not only objects, but also the events in which these objects are involved.

At one level of comprehension, our mind represents objects as "things that are in the world", such as birds and apples and cups. Once we add a temporal level of comprehension, and thus we observe how things change or preserve their own identity through time, we also keep track of what causes may change the properties of an object. The focus of EST is on events, which are treated as 'pegs', basic computational units or 'slots' on which we stack up information, and which stand for relations and order among relations in which objects are involved, as they unfold over time (Speer *et al.* 2003, Zacks 2004, Zacks *et al.* 2007, Zacks & Swallow 2007, Reynolds *et al.* 2007, Tversky *et al.* 2008).

EST assumes that, at a minimum, we can observe objects at two levels. One basic level is that of their structure and how it is realized in space (*partonomy*) and one of an object and its relation to an abstract class (e.g., a chair as part of furniture: *taxonomy*). Once we take in consideration a temporal dimension, in which objects can have different properties in different intervals of time, we will have 'dynamic' objects or *events*. Events are conceived as discrete information units derived (i.e. *transduced*) from perceptual information, i.e. the 'indexes' attributed to the combination of a (rather abstract) visual property and the object that instantiates it.

For instance, if someone throws a cup on the floor, then the cup will likely be shattered into pieces because of this action. The temporary relation between an individual and the cup will bring about a new state of affairs in which the cup will be a new property of some sort, that of being shattered. At the same time, we represent this change via the *temporal* and *causal* relation between the two state of affairs, one involving an event of someone shattering the cup, and another in which the cup will be shattered, which is separated by a boundary event, an interval of time in which neither the cup is shattered nor it is still intact, and in which we will need to 'update' our model of events. Events can also be combined together: If someone is stacking pillows, each single pillow-stacking event can be combined into a 'bigger' pillow-stacking event, and possibly 'merged' with other events, forming a more complex event such as 'pillow-ordering'.

Such complex sequences of events can be seen as event models or schemata, structures of events and their causal/temporal connections, as they are represented in *short-term memory* (models) or stored in *long-term memory* (schemata in 'semantic memory': see e.g. Tulving 1972, 1983, 2000a, 2000b, 2002 and references therein for an introduction). Events can be dynamically bound: The "throwing" event occurs at a time *t+1*, a boundary event is formed at a time *t*+*2* and the "shattering" event occurs at a time *t+3*, then there will be a causal, as well as temporal relation between these events.

Both MOT and EST are theories that offer a detailed picture of how dynamic vision can occur, defining in detail the mechanisms by which we track objects in motion, and the complex spatial representations that arise from this process, or events. One alternative view to these approaches that offers some further important insights on spatial representations is the *Hippocampus as a Cognitive Map* theory (HCM) of O'Keefe & Nadel (1978). HCM started as a study of rats' navigational system, the way they represent objects and their places in the environment, and how this information is memorized and accessed or updated at later stages. According to this theory, humans (and rats) build up a complex spatial representation of the environment via two parallel systems: the *place* and the *misplace* system. The place system records information about objects' position in the environment, and 'checks' whether this information is correct when visual information is processed. If an object has changed position, then the misplace system records the change of position and updates the object's new position accordingly.

This model has been further extended over the years. O'Keefe (1983, 1990, 1991) and Burgess & O'Keefe (1996, 2003) show that information about objects and their relations is processed, in 'real time', by the navigational system. This system computes the location of a figure in terms of polar angle Φ and distance d from the ground as the relation $\theta(\Phi,d)$, computed via *θ-rhythm* signals, which mostly originate in the Hippocampus.

The result of these computations can be modeled as a vector, a sequence of cells (*Boundary Vector Cells*) that fire when an observer visually tracks relevant entities in an environment, and can also allow to compute geometrical properties of objects. Hence, the place and misplace systems build up complex spatial representations over time, or Cognitive Maps (O'Keefe & Burgess 1999, Burgess *et al.* 2002, and Burgess 2006a, 2006b; see Arsenijević 2008 for a linguistic proposal).

These theories give us a basic insight on the nature of dynamic spatial vision. When we keep track of objects in motion, we do via the properties that objects may have over time, whether they are geometrical, functional or 'functional', insofar as they allow us to track objects in space. At the same time, we also keep track of the relations between these properties and their order of causal/ temporal occurrence: Spatial representations have an inherent temporal dimension, which represents the structural relations between the events making up these representations.

Adpositions, as the chief part of speech expressing these relations, must also have such an abstract nature. Look at the examples:

(18) Mario has fallen onto the floor.

(19) Mario has gone into the room.

(20) Mario is sitting near the patio.

A scenario which is more or less depicted by (18) and (19) is one in which Mario is on the floor and in the room, respectively, as a consequence of a particular event, one of falling and one of going. A scenario depicted by (20) is one in which Mario is involved in an event of sitting, which occurs at some distance from the patio. He may be involved in other events, although these events are in a sense 'backgrounded', in the sentence.

In all three cases, the spatial relation holding between Mario and different grounds holds at some moment of time because of some previous event, and involves more than just geometrical information. If we conceive Mario and the floor as inherently visual objects, then the adposition *onto* will capture not only that these two objects are currently related one another via a 'support' relation, but also that such relation has come into being because of a previous falling event. Since adpositions seem to express the 'logical' structure behind the events described by a sentence, the kind of spatial representations they capture are representations in logical space, and define possible relations between objects and how they are represented in this logical space. I shall offer the precise logical details of this enriched logical space in the next section, in the proposal I shall call Visual Representation Theory.

3.4. *A Logic of Vision, Part II: A Model of Visual Logical Space*

In the previous section, we have been able to define a richer notion of visual object (i.e. things and their spatio-temporal properties), as well as sketching the nature of the relations holding between these objects. I shall integrate these results in our logic of vision as follows.

First, I shall assume that the set V of visual objects is now made of 'structured' entities, the combination of events e, objects o, and properties pr. The complex visual object that is made of these elements is the triple $v=<e,o,pr>$, a basic entity which I shall call *Visual Representation Structure (VRS)*. Importantly, the set of events E (with $e\leq E$) is disjointed from that of objects O (with $o\leq O$) and the union of the two sets forms the whole set of (basic) visual objects, i.e. $E\cap O=0$ and $E\cup O=V$. Properties pr form a set of properties by which these visual objects can be individuated, i.e. we have $pr\leq PR$. In words, VRSs are made of basic objects (e.g. 'x'), the properties by which we individuate them (e.g. "swan"), and the events in which these properties are instantiated, i.e. their position in logical space with respect to other events. The following will hold: $v\leq V$, i.e. each VRS is part of the set of VRSs. I shall represent a VRS as $e{:}pr(o)$, which reads: An event e instantiates a property pr of an object o. This format follows the format of DRT, which I shall introduce in full in section 3.5.

We thus have seen that VRSs can be combined together via *merge*. The sum of two VRSs can be seen as a complex, novel event in which different properties of the same object can be combined together into a more complex, novel property. If $e{:}grab(x)$ and $h{:}order(x)$ are respectively an event of (pillow) grabbing and (pillow) ordering, then the complex event of (pillow) clean up can be formally defined as: $(e{:}grab(x)+h{:}order(x))=i{:}clean\text{-}up(x))$.

The structural properties of *merge* (associativity, commutativity, idempotence) are defined over VRSs as well, although the apparent 'temporal' nature of VRSs as representing 'objects in motion' requires some discussion. I shall focus on events, to make the discussion simple. An event of (pillow) clean-up can be organized in different ways (associativity); while we usually first grab pillows and then order them, when we clean up, an event of (pillow) clean-up consists of both events, regardless of their linear order (commutavity); several events of pillow-grabbing are still a (complex) event of pillow-grabbing (idempotence). Although VRSs are more complex objects, their 'combination' can be nevertheless defined via one basic operation, that of *merge*, which represents how complex VRSs are created from the union of basic VRSs.

The *part-of* relation is also defined over VRSs and events, and allows to define how events are structured. Reflexivity and transitivity allow to establish order/overlap among complex sequences of VRSs, straightforwardly enough. Antisymmetry allows to establish whether two VRSs (or parts thereof) are really the same, and thus to establish the identity between a complex VRS and the sum of its constituting VRSs. It also allows us to reconstruct their consequential/ temporal relation as well: if $e{:}grab(x)\leq i{:}clean\text{-}up(x)$, then $e{:}grab(x)\cup i{:}clean\text{-}up(x)=i{:}clean\text{-}up(x)$ and $e{:}grab(x)\cap i{:}clean\text{-}up(x)=e{:}grab(x)$. Since an event of pillow-grabbing is a proper part of (pillow) cleaning up, then it must precede the realization of a cleaning up event. The structural relations between events thus represent their causal/temporal relations: 'New' events come into being as the result of 'old' events being combined together, in an incremental fashion. If we don't grab and order pillows, we won't have an event of pillow-cleaning up: The existence of this event is a consequence of the combination of previous events.

The tight relation between the syntax and semantics of our logic of vision thus allows us to capture one aspect of 'dynamic' space by simply looking at how events are computed, without introducing further principles of analysis. Our new logic of vision can be thus represented as $S=<V,+,\leq>$, with V being a shorthand for $V=<E,O,PR>$.

This new logic of vision is a fully dynamic logic of vision when combined with an index set I, i.e. when we have $S_d=<I,V>$, with I being the index structure $I=<t,+>$. It allows us to explicitly represent how we integrate VRS together, one example being the following:

(21) $t.$ $e:grab(x)$ (VRS instantiation)
$t+1.$ $h:order(x)$ (VRS instantiation)
$t+2.$ $e:grab(x)+h:order(x)$ (*merge* introduction)
$t+3.$ $e:grab(x)+h:order(x) =i:clean\text{-}up(x)$ (sum of events)
$t+4.$ $e:grab(x)\leq i:clean\text{-}up(x)$ (*part-of* rel. intr.)

In words, the merging of two VRSs yields a more complex VRS as the result, and allows to establish structural relations between VRSs. As we can see, the use of dynamic binding also allows us to bring out one aspect of the temporal nature of events: If we grab a pillow at a time t and then put it in order at a time $t+1$, then the resulting pillow-cleaning up event will be realized as a later time $t+3$, in a progressive way.

At this point, we have a quite rich and thorough *logic of vision* which allows us to model spatial representations/relations in a rather elegant and simple way, and which turns out to be somewhat similar to similar other logical theories proposed in, for example, the AI literature (e.g. *Event Calculus*; see Hamm & van Lambalgen 2005 and references therein and van der Does & van Lambalgen 2000 and e.g. Barwise & Seligman 1997 for non-linguistic applications of situation semantics). One example of the elegance behind our logic is the notion of 'location'. VRSs explicitly represent the spatio-temporal 'location' of some event and its participants by representing the properties that individuate these entities. Geometric or mechanical properties are not any different from 'grabbing' properties, with respect to how this process occurs over time: we can thus represent e.g., the notion of inclusion as the VRS $e:in(x)$, that of support as $e:on(x)$ and so on.

We can then represent the notion of 'motion', or more appropriately the notion of change, as an ordering (*part-of*) relation between VRSs and the events they represent. So, if Mario goes in direction of the room and then stops once he's in the room, he will be in the room as a consequence of this event of motion. This can be represented as $e:go(r)<i:in(r)$, i.e. an event of going into the room as expressing the relation holding between one event and its consequence. In a scenario in which Mario is sitting near the patio instead, other events may be going on at the same time, but at least these properties allow us to individuate Mario. We can represent this as $n:near(p)\leq z:gen(p)$, an event of (sitting) near the patio as part of a more generic event.

These relations between VRSs and the events they represent may find their way into language chiefly as adpositions via the function f, the isomorphism between vision and language. I shall re-define this function as follows. If f takes a pair of visual object and property as an input, it will return a noun as an output — we have $f(<o,pr>)=\mathbf{n'}$. If *f* takes a pair of event and property as an input, it will return a verb as an output — we have $f(<e,pr>)=\mathbf{v'}$**.** If it takes a full *VRS* as an input, it will return an adposition as a result — we have $f(<e,o,pr>)=\mathbf{p'}$**.**

The intuition is that 'partial' VRSs find their way in language as (common) nouns, labels for objects, and as verbs, labels for 'actions'; both individuate some entities, but do not express relations between these entities. Adpositions instead express the structural relations between VRSs, ultimately complex VRSs. The intuition is simple: Nouns (and verbs) find objects in logical space; adpositions denote the relations between these objects, which in turn represent a very abstract notion of space. L&J's syntactic proposals are still maintained, to an extent.[12]

The function *f*, as an isomorphism, preserves structure on VRSs as well: An adposition like *between*, for instance, is usually analyzed as the 'sum' of two simpler adpositions, such as *to the left or to the right of* some ground (e.g. Zwarts & Winter 2000). This can be represented as $f(r\text{-}of+l\text{-}of)=f(r\text{-}of)+f(l\text{-}of)$, i.e. the adposition representing the "between" relation is lexically equivalent with the adpositions representing the relations "to the left of" and "to the right of". Generalizing a bit, from basic spatial representations we can build more complex spatial relations; the complex structure defined by this process, the model of logical space defined by our logic of vision, may be represented in language *up to isomorphism*, via process of progressive refinement and specificity of relations (cf. also Levinson & Meira 2003). Hence, the mapping function *f* may assign different labels to its outputs, depending on the level of fine-grainedness and with some consistent cross-linguistic variation, but in a quite fine-grained and structurally regular way (cf. again Talmy 2000; see also section 4.2). Again, via this function we represent the possibility that we can match for each VRS a corresponding linguistic unit, and that the structural or 'spatial' relations between VRSs can find their way into language, chiefly as adpositions, at least in a language such as English.

Before moving to language, however, I shall make one observation regarding the nature of this process. According to the HCM proposal, when we mentally represent visual objects, these objects can be seen as output to some previous visual, perceptual input, which is then transduced as a visual object. This process occurs over discrete intervals of time, which in turn may be seen as minimal cycles of the *θ-rhythm*, and which may actually occur *independently* of the presence of external stimuli. In the absence of external stimuli, our brain still partitions the perceptual stream into minimal, discrete units. Very informally, our vision faculty will organize the perceptual stream into minimal units even if we are not observing finite objects such as tables, or if we look at the same portion of sky for quite a long interval of time.

When external stimuli are tracked, then it is possible to check whether they stand for some 'new' or 'old' information, i.e. whether their internal representation matches previous visual computations. Hence, the underlying properties

[12] In this paper, I shall not propose an explanation on *why* the function *f* seems to operate such distinctions in the labeling process, and leave such a complex topic for future research.

of these computations do not crucially hinge on external stimuli, but on the possibility (perhaps, necessity) to integrate these different forms of information together in an effortless way, and in a coherent, 'synchronized' model (e.g., O'Keefe 2003, Buzśaki 2006). Our logic of vision thus represents an internal model of logical space, and represents the properties and relations defined over this model. By this point, our discussion of the logic of vision should be thorough enough: I shall concentrate on spatial language and its logic.

3.5. A Logic of Language: Discourse Representation Theory and Space

The study of meaning in natural language as a psychological phenomenon has long been adversed in model-theoretic approaches, traditionally rooted in an 'anti-psychologist' philosophy (e.g. Davidson 1967, Montague 1973, Cresswell 1985). Some modern research, however, broke with this tradition and attempted to study whether the models defined in this approach can be seen as mental structures and processes of some sort, represented via *dynamic logic* (e.g. Kamp 1981, Heim 1982, Chierchia & Turner 1988, Kamp & Reyle 1993, Chierchia 1995).

Among these different approaches, *Discourse Representation Theory* (DRT) represents the most important theory with such a 'cognitive' stance, and offers a set of tools which will allow us to (easily) treat all the linguistic phenomena we shall address via a single set of formal tools. For instance, it includes detailed treatments of the semantics of nouns and temporal expressions, which can be extended to treat our adpositions data (e.g., theories of noun reference and plurality such as Link 1983, 1998 or treatments of events such as Parsons 1990 and Landman 2000, 2004). It also allows us to take a perspective to sentence interpretation as a dynamic process, since it aims to model how sentences are interpreted and 'used' to form models in a compositional and incremental and on-line fashion, as in models of parsing such as Crain & Steedman (1985).

The version I shall use here is also fully compositional and thus allows us to analyze the contribution of each word to a sentence (iKamp *et al.* 2005, based on Muskens 1996 and van Eijck & Kamp 1997), and may be ideally implemented with certain minimalist theories of syntax with a 'processing stance' (parser-is-grammar of Phillips 1996). However, I shall focus on the contribution of nouns and adpositions for the most part, being somewhat sloppy on other parts of speech (such as verbs). Although the structural equivalences with my logic of vision should be immediately obvious, I will defer a thorough discussion to section 4.1, and focus here on the linguistic bits.

The most basic bits of information in DRT are Discourse Representation Structures (DRSs). A DRS can be thought, at a minimum, as a linguistic information state containing a set of discourse referents (or *U* for universe), an 'object' in discourse, and the conditions (or *CON*) which allow us to individuate such objects in discourse. While basic ('extensional') DRSs are at a minimum a duple of discourse referents (or individuals, for the sake of clarity) and their associated conditions, they 'become' information states when a third set of objects is taken in consideration, possible worlds (the set *W*). Hence, a DRS or information state is the triple *<W,U,CON>* or *<{w},{x},{**con'**(x)}>*, in which a discourse referent is paired with a 'world' referent and a condition, and which can be seen as a mental representation or (mini-)model that a speaker entertains, when he parses chunks of sentences, incrementally.

The nature of this world 'coordinate' deserves a few words of discussion. In classical logic, possible worlds are seen as quite real Leibnizian entities, such as the world we live in (e.g. Lewis 1986). Many versions of DRT, however, propose a different approach, partially based on Stalnaker's (1973, 1999) work,[13] in which possible worlds are mental objects, and represent nothing else than possible scenarios in which referents are involved, those for instance expressed by a sentence or a more complex text. Consequently, possible worlds can vary in 'size' and structure, and may be intuitively related one another according to the same principles definable over individuals, DRSs or other model-theoretical objects, as assumed in situations semantics (e.g., Barwise & Etchemendy 1990, Kratzer 1989, 2007) or modern modal logic (Hughes & Cresswell 1996, Blackburn *et al.* 2006).

Let us now turn to formal matters. As a standard convention, I write conditions in boldfaced characters and by adding a prime, i.e. '**con**''. Hence, conditions in DRT are roughly equivalent to non-logical constants of first-order logic, and thus they represent 'concepts' or 'thoughts' as they are expressed in natural language, together with the distinction between *intension* and *extension* (cf. Margolis & Laurence 1999, Gärdenfors 2000, Winter 2008). The obvious consequence of this assumption is that our concepts/conditions will thus be invariably complex and definable in terms of their internal structure, unlike assumed in atomistic theories of concepts such as Fodor (1998, 2003). While an interesting topic per se, its discussion would lead us too far afield from our main topic of discussion, so I shall leave it aside for the time being.[14]

For our purposes, worlds and eventualities (i.e. events, properties changing over time, and states, properties holding over time) are basically the same (model-theoretic) objects, as in some variants of situation semantics. Very informally, if individuals represent objects, then eventualities represent the relations in which individuals are involved.[15] I shall use the term 'events' and avoid making any distinction between events and states, for the sake of clarity.

Once I have defined the basic structures of DRSs, I shall focus on the combinatorial and interpretative apparatus, i.e. how DRSs can be used to represent linguistic expressions. Here I shall use a variant of the 'linear' notation, rather than the more popular 'box' format, to enhance readability (as in Geurts 1999). I shall roughly match one syntactic phrase with one *DRS*, although more precise analyses are possible (see Kamp *et al.* 2005 for discussion). Look at the example:

[13] This is true insofar as we look at the 'raw mechanics' of the underlying logic. Stalnaker's position is not a mentalist/internalist one: For him, 'possible worlds' are those of classical logic. DRT offers a much stronger mentalist perspective. Very informally, 'worlds' in DRT are roughly equivalent to possible thoughts or beliefs, information states ascribed to (thinking) agents. See Maier (2006: chap. 1) for discussion.

[14] I would like to thank an anonymous reviewer for bringing this topic to my attention.

[15] Note that, informally speaking, events and states are included in intervals of time, within the DRT architecture, with intervals of time forming up the main 'temporal structure' of a discourse. I shall diverge from DRT and use intervals of time in a different way, as I shall show in the remainder of the section.

(22) A man walks quickly. He whistles.

When a sentence like (22) is parsed, the parser builds up a bottom-up, left-to-right syntactic representation and, for each constituent and phrase, it builds up the corresponding DRS. For instance, *a man* is parsed as noun phrase/determiner phrase, and interpreted as the DRS [x:**man′**(x)], a DRS representing a referent x and a condition individuating him.

The next step consists in combining the predicate *walks* with the noun phrase *a man*. This is obtained via the syntactic operation *merge*, which shall represent as '+'.[16] *Merge* in DRT is a binary (associative, commutative, idempotent) operation that takes two DRSs and gives a 'bigger' (or new) DRS as the output, by unifying the universes and conditions of each DRS. In more formal terms, we have:

(23) [{x}:**con′**(x)]+[{y}:**con′**(y)]=[{x,y}:**con′**(x),**con′**(y)] (*merge* introduction)

In words, the merging of two DRSs forms a bigger DRS in which the universes and the conditions are merged pair-wise. Merged conditions are interpreted as being conjoined. If we were to translate conditions from our DRT language to first order logic, merged conditions would be interpreted as being conjoined, whereas each referent in the universe of discourse can be translated as an existentially quantified variable. We would have "∃x∃y[**con′**(x)&**con′**(y)]" for the two conditions in (23) (cf. Kamp *et al.* 2005: 143–145). I shall use brackets to mark the universe, and thus enhance readability (e.g. *{x,y}*), as in van Eijck & Kamp 1997 and Kamp *et al.* (2005).

The verb *walks* can now be simply represented as [e:**walk′**(x)], i.e. a DRS which introduces no new (object) referents but a novel spatio-temporal referent, the event of walking. The merging of the two resulting DRS can be represented, in a piece-meal fashion, as:

(24) *t.* [{x}:**man′**(x)]+[{e}:e:**walk′**(x)]=[{e,x}:**man′**(x), e:**walk′**(x)] (*merge* intr.)
t+1. [{e,x}:**man′**(x),e:**walk′**(x)]+[{e}:**quickly′**(e)]= (*merge* intr.)
t+2. [{e,x}:**man′**(x),e:**walk′**(x), e:**quickly′**(e)]

In words, we obtain the DRS representing the first sentence in (22) (*A man walks quickly*), by merging the DRSs representing its main constituting phrases. The DRS for *a man* acts as the context DRS, which is then updated via *merge* by the DRS for *walks*, acting as the context change potential DRS. The dynamic aspect of meaning is thus represented by the ability for new phrases/words to add more information regarding referents and events represented by each sentence, and thus define a 'broader' model representing facts. This is also represented via the explicit use of an index set in the derivations, which allows to explicit represent

[16] Kamp *et al.* (2005) use a different symbol, but this difference is immaterial, for our purposes. Note also that the properties of merge (associativity, commutativity, idempotence) stem from its definition as a (complex) form of set union, with idempotence allowing to 'reduce' universes whenever they are identical (see e.g. (22), i.e. *{e,x}+{e,x}={e,x}*).

how DRSs are combined together (as in Muskens 1996 and van Eijck & Kamp 1997, for example).[17]

The merging of DRSs has also one important consequence: it defines a semantic level of relations between DRSs and their universes/conditions, the *accessibility/part-of* relation. The *accessibility/part-of* relation is a transitive, anti-symmetric, reflective relation which allows to define one DRS *d* as part of another DRS *d′*, i.e. *d≤d′*. While transitivity and reflexivity intuitively define how DRSs are connected over the flow of discourse, antisymmetry allows to make establish what relation holds between two referents/events/DRSs. One example is pronoun resolution: Intuitively, a pronoun such as *he* in (20) denotes one whistling individual as being a specific individual out of those who are walking quickly in the park. If at least part of the content expressed by two DRSs can be the same, then the two DRSs individuate the same object, a condition which expresses an anaphoric relation and is usually represented as *x=y*.[18] When the *accessibility* relation is restrained to discourse referents or events, it is usually called *part-of* relation (e.g. Kamp *et al.* 2005: 135). Consequently, I shall just use the *part-of* label for a semantic relation holding between DRSs, in order to make the exposition of the arguments clearer.

This is shown in the remainder of the derivation for (20):

(25) *t+3.* [{e,x}:**man′**(x),e:**walk′**(x),e:**quickly′**(x)]+[{e,y}:y=?,e:**whistle′**(y)]=
t+4. [{e,x,y}:**man′**(x),e:**walk′**(x),e:**quickly′**(e),y=x,e:**whistle′**(y)]

In words, the merging of the first and second sentence will also establish an identity relation between first walking man and second whistling man: There is really one man we are talking about, in (22). The resolution of the open anaphoric relation (i.e. *x=?*) amounts to identifying two referents by stating that the properties by which these referents are individuated converge to the same result.

After this brief introduction to the relevant aspect of DRT, I shall focus on a compact treatment of adpositions, which diverges from the standard DRT treatment of this category (cf. Kamp *et al.* 2005: chap. 2–3) and introduce a more thorough analysis of these terms, based on the vast literature on the topic. My basic assumption will match the non-linguistic considerations I offered in the previous section: Adpositions denote relations between DRSs, by expressing how the events denoted by these relations are ordered (e.g., Kamp 1979a, 1979b, Jackendoff 1983, Parsons 1990, Wunderlich 1991, Nam 1995, Fong 1997, Kracht 2002, Landman 2004, Zwarts 2005, Svenonius 2006, Ramchand 2008, and Kratzer, to appear).

[17] In the dynamic semantics literature, the notion of ‘dynamic binding’ has a more restricted (semantic) application, and it is restricted to inter-sentential *merge*, i.e. the binding of information units over the sentence boundary (e.g., Chierchia 1995, Stockhof *et al.* 1996).

[18] Pronoun resolution is sensible to features, like gender and number or temporal/aspectual values. I just ignore these aspects here, for the sake of clarity. In DRT, pronoun resolution also involves *presupposition* resolution, what could be (very) informally defined as the integration of implicit information in a DRS, together with the resolution of the anaphoric relations associated with this implicit information. See, among others, van der Sandt (1988, 1992), Geurts (1999), and Kamp *et al.* (2005: chap. 1–2) for discussion and references on this very complex and rich topic.

I shall thus assume that adpositions denote anaphoric relations between events/DRSs. Differently from pronoun anaphora, though, they may express 'asymmetric' relations, i.e. relations in which events are not necessarily identical. In this perspective, adpositions are akin to the 'duplex conditions' of DRT, which are used to represent quantifiers such as *every*, but also conditionals (e.g., donkey sentences), temporal adverbs and other temporal/logical relations.

The main reason for this assumption can be motivated by the following entailment patterns in the examples (adapted from Parsons 1990):

(26) A delegate walked into the park. → A delegate was in the park.

(27) A delegate is near the park. → A delegate is near the park.

In (26), the sentence *A delegate…* entails that the relevant delegate was in the park as a consequence of this event of motion. In (27), the sentence *A delegate…* entails itself, in the sense that it the delegate's position is not an explicit consequence of some previous event of motion, but also holds for possibly more specific states (e.g., the delegate being currently near the park). The symbol '→' represents the entailment relation between the two pairs of sentences.

The intuition behind these patterns is simple: adpositions, as they mirror relations between VRSs in language, also denote equivalent relations between DRSs and the events included in these DRSs. They do so by explicitly stating how events are ordered one another, thus explicitly representing the causal/temporal structure of (parts of) a sentence, possibly restricting this relation to certain events (e.g., those being "in" the park). I shall thus translate *into* as the complex DRS [{e,s,x,y}:*e<s*,s:**in'**(x,y)] and *near* as the complex DRS [{e,s,x,y}:*e≤s*,e: **near'**(x,y)]. The *DRSs* represent in a compact manner the Parsonian entailments, as *part-of* relations between the events denoted by the merged sentences. Informally, if a delegate walked into the park, he was in the park as a consequence. If a delegate is near the park, he may have arrived there because of some other events, or may stay there for some unspecified interval of time.

The interpretation of (24), at the relevant step and abstracting away from tense, is the following:

(28) *t.* [{e,x}:**delegate'**(x),e:**walk'**(x)]+[{e,s,y}: *e<s*,s:**in'**(x,y),**park'**(y)]=
t+1. [{e,s,x,y}:**delegate'**(x),e:**walk'**(x),*e<s*,s:**in'**(x,y),**park'**(y)]

In words, (28) says that a delegate walked and, as a consequence of this event of walking, he ended up in the park. The interpretation of (23) would be similar, except that the contribution of "near" would yield the following (slightly informal) *DRS*: [{e,s,x,y}:**delegate'**(x),s:**be'**(x),*s≤s'*,s:**near'**(x,y),**park'**(y)].

This treatment of English adpositions is by no means exhaustive and would probably need revisions, especially once we take in account a broader cross-linguistic perspective and the well-known interplay of adpositions and verbs of motion (again, see e.g. Talmy 1978, 2000, Svenonius 2006, Higginbotham 2009, and Zwarts 2010 for discussion). However, it allows us to represent in a rather simple what kind of contribution adpositions (and nouns) offer to a sentence, as well as introducing a rather compact theory of linguistic representation, in the guise of DRT. I shall thus collect all the crucial aspects of DRT and present them as parts of DRT's underlying logic.

DRT can be treated as a logic of language, which can be represented as the model L=<D,+,≤>. The set D of DRSs is in turn a set of triples, defined as d=<w,u,***con***>, and with d≤D holding for each d. The *model* (of discourse) defined by DRT is a *lattice* which has a structure entirely equivalent to that defined for vision.[19] The 'dynamic' incarnation of this model is L_d=<I,D>, the duple formed by DRSs and intervals of time at which they are combined together, with I again being defined as I=<t,+>.

The mapping from this model of language to other models, most specifically our logic of vision, can be easily defined via the function g, which is usually known as the *anchor function* in DRT (Kamp *et al.* 2005: chap. 4, Maier 2006: chap. 3 for discussion). This function is defined as an isomorphism mapping each linguistic information unit onto a non-linguistic unit, in this case a visual unit, i.e. $g(\boldsymbol{d'})=v$: In our case, it matches DRSs (linguistic information units) with VRSs (non-linguistic, visual information units).

Since it is an isomorphism, it maps at least one DRS onto one VRS, and at most one DRS onto one VRS. It preserves structure, so a mini-discourse like (22) can be seen as the description of a complex scenario, made of two connected, simpler scenarios. Formally, we have $g(\boldsymbol{d'}+\boldsymbol{k'})=g(\boldsymbol{d'})+g(\boldsymbol{k'})$, which in words says: The scenario corresponding to the mini-discourse in (22) corresponds to the scenario matched by the first sentence (a man is walking in the park) followed by the scenario matched by the second sentence (this man is whistling). Much like the function f, the function g can, but needs not to, find a VRS for each mapped term. In this regard, the function g can also be thought as representing a top-down process, since it represents how we can consciously match a sentence (and its content) with an extra-linguistic scenario it refers to.

Now that both sides of the isomorphism are defined, we have a good understanding of how information flows from vision to language and from language to vision, and thus we are ready to tackle the problem of the vision–language interface in an explicit way. However, before doing so, I shall offer an answer to the second research question, which is now within our reach. The answer is the following:

A–2: *Our models of spatial vision and language must include any possible property and relation that can 'connect' two entities; these models can (must) be treated via a model-theoretic approach.*

Spatial vision and language, then, can be seen as two systems representing different aspects of the same underlying phenomena: How we build up and maintain complex 'maps' of the objects we keep track of, over discourse. At this point, we can explore the common space generated by these two structures, and

[19] In DRT or similar approaches (e.g., Krifka 1998), events and referents are part of (structurally) different structures; here I follow Link (1983, 1998) and assume one common type of structure for all types of object.

thus focus on the vision–language interface.

4. A Theory of the Vision–Language Interface, and Beyond

In this section I shall offer a logical theory of the vision–language interface based on the results of the previous section (section 4.1); I shall offer empirical evidence in support of this approach (section 4.2); and sketch some broad consequences of my approach with respect to theories of the language faculty (section 4.3).

4.1. The Vision–Language Interface: A Formal Approach

A theory of the vision–language interface, given the discussion so far, must be a theory about the two-way information flow between two structures which represent (external) spatial information in a principled and highly organized way, the logical space defined by the logic of vision and language. As section 3 constituted a relatively long analysis of how these notions emerge from the basic bits of vision and language, I shall re-state my basic assumptions first and then dive into the vision–language interface problem.

I have assumed that both vision and language can be represented via a precise logic, which I called the logic of vision and the logic of language, respectively — or Visual Representation Theory (VRT) and Discourse Representation Theory (DRT), equivalently. These logical calculi share the same underlying structure: VRT is defined as triple $S=<V,+,\leq>$ and DRT as the triple $L=<D,+,\leq>$. These models are lattices, partially ordered sets, which minimally differ in having different types of elements, rather than in their structure.

The basic elements in these domains are respectively VRSs and DRSs: for each VRS v, the relation $v\leq V$ holds; for each DRS d, the relation $d\leq D$ holds. For each VRS v, the following identity holds: $v=<e,o,pr>$, i.e. each *VRS* is a triple of an event, an object and a property that identifies an object in an event. For each DRS d, the following identity holds: $d=<w,u,con>$, i.e. each DRS is a triple of a world/event, a referent and a condition that identifies a referent in a world/event. While VRSs are discrete units (possibly) representing perceptual stimuli from the visual apparatus, via transduction, DRSs may be seen as discrete units representing other types of information units (e.g., 'concepts' or 'thoughts'). They may be connected to VRSs via a slightly different type of transduction, but do not have a direct 'external' grounding: They represent purely 'internal' information.

While the two structures have different elements, their operations and relations are basically the same. A syntactic operation, *merge* (ultimately, set union), allows to define each element as the sum of other elements, possibly only itself. We represent it as '+'. Its definition is simple: It is a binary operation taking two inputs of the same type (e.g., DRSs: $a+b$), yielding an output of the same type as the inputs (a DRS: $a+b=c$). It is *associative*, *commutative*, and *idempotent*: It allows to combine the same elements in different ways (associativity: $(a+b)+c=a+(b+c)$), regardless of their order of occurrence (commutativity: $a+b=b+a$), and can be 'repeated' on the same input (idempotence: $a+a=a$).

A semantic relation, the *accessibility/part-of* relation (represented as '≤'), integrates this syntactic operation and establishes how the results of the *merge* operation are 'connected'. It is binary, as it establishes a relation between two objects of the same type (e.g., VRSs: $a \leq b$, and it is reflexive, asymmetric and transitive: It allows us to establish that objects are part of themselves (i.e. $a \leq a$), that objects can be identified (i.e. if $a \leq b$ and $b \leq a$, then $a=b$), and that multiple relations can be compressed as a single relation (i.e. if $a \leq b$ and $b \leq c$, then $a \leq c$).

The *merge* operation and the *part-of* relation are connected via the following properties, which I shall again represent via set-theoretic notation. If $a \leq b$, then $a \cup b=b$ and $a \cap b=a$. In words, if one object is part of another, then their merging will correspond with the 'bigger' object (union), and their product will correspond to the 'smaller' object (intersection). Semantic relations can be seen as the result of previous instances of syntactic operations, in a sense recording the successful merge of two objects into a more complex, novel object. The structures defined by these operations are complex Lattices, i.e. partially ordered sets with a syntax and a corresponding semantics, and thus models of the phenomena they represent.

Although other operations can be defined (e.g., set intersection standing for attention), this 'minimal' logic allows us to aptly model how information units are processed and integrated together into more complex units, in a bottom-up way. They also allow us to define how one logic can be tightly connected to another via two functions, f and g, which respectively define an isomorphic mapping from VRSs to DRSs and from DRSs to VRSs. These functions are isomorphic because they map at least one input and at most one input to the same output, i.e. they are respectively injective and surjective, thus they are bijective.

The function f is defined as: $f{:}v \rightarrow d$, i.e. a function that maps each visual structure $v \leq V$ onto a discourse structure $d \leq D$, whereas the function g is defined as: $g{:}d \rightarrow v$, i.e. a function that maps each discourse structure $d \leq D$ onto a visual structure $v \leq V$. Note, now, that these two functions are one the inverse of the other: Their *composition* (represented via the symbol '$\circ$') will yield the identity function, e.g., we have $f \circ g=i$, with 'i' being the *identity function*. This latter property tells us that, for example, each noun may act as the linguistic label for a visual object, and thus that each visual object may a have noun as a linguistic label.

These isomorphisms allow us to explicitly represent how we 'translate' one type of objects into another, while for logical operators (i.e. *merge* and the *part-of* relation), they offer evidence that these operations are the same across models/ logical systems. The reason is simple: while objects define non-logical constants, *merge and the part-of* relation define logical constants, elements of a logic that receive the same interpretation on any model, whether it represents vision or language. In words, *merge* is interpreted as the union of two objects, whether these sets stand for visual structures or discourse structures, and so is the *part-of* relation interpreted as a relation between objects.

This is explicitly represented via a structure-preserving condition on our isomorphisms: $f(a+b)=\boldsymbol{a'}+\boldsymbol{b'}$, given that $f(a)=\boldsymbol{a'}$ and $f(b)=\boldsymbol{b'}$. In words, the noun for the object corresponding to the merge of the object "legs" and "surface" (*table*) corresponds to the super-ordinate noun that stands for the objects "legs" and "surface". The *merge* symbol is the same on both sides of the identity, while the merged objects are different. The same holds for the *part-of* relation, since, if we have $a\leq b$, we have $f(a)\leq f(b)$. In words, if a leg is part of a table, then the noun/concept *leg* is part of the noun/concept *table*. The same considerations hold, *mutatis mutandis*, for the function g. In words, vision and language may differ as models representing different 'things', but they are equivalent as models sharing the same structure.

The definition of these two isomorphisms has one important consequence: It allows us to outline a simple and yet very precise theory of the vision–language interface. The main assumption I shall make is that the vision–language interface is defined as a *Galois connection* between these two structures. A Galois connection is defined as follows: given two lattices $<A,\leq>$ and $<B,\leq>$, $f(a)\leq b$ *if and only if* $a\leq g(b)$. In our case, and with some notational fantasy, given the lattices $<D,\leq>$ and $<V,\leq>$, we have $g(\mathbf{d'})\leq v$ if and only if $\mathbf{d'}\leq f(v)$. In words, if vision and language are connected via a Galois connection, then the VRS corresponding to a DRS is part of a larger VRS, and a DRS corresponding to a VRS is part of a larger DRS. In words, vision and language representations are connected if each linguistic term is matched by a visual entity, which is part of a 'larger' scenario, and if each linguistic term expressing a visual object is part of a sentence. Informally, a Galois connection is a method of defining an isomorphism between structures in which weaker relations can also be defined: it allows us to express not only that structures 'look the same', but also to compare the relation between many elements of one structure to an element of the other structure (e.g., Ganter & Wille 1998: chap. 1).

The strength of this proposal is that it allows us to define a degree of accuracy by which a certain sentence describes a state of affairs and *vice versa*. For instance, an adposition matches a spatial representation when the two following conditions hold: $f(v)=\mathbf{d'}$ and $g(\mathbf{d'})=v$. In words, if a book is supported by the top vertical surface of a computer, then the adposition *on top of* is quite ideal match for this scenario, since we intuitively have $f(\textit{on-top})=\textbf{on-top'}$, but also $g(\textbf{on-top'})=\textit{on-top}$.

While identity cases are in a sense trivial, cases of partial matches allow us to grasp the crucial strength of the proposal. For instance, an adposition expressing only support of the book by the computer is intuitively less accurate (i.e. *on*) than *on top of*, which expresses the specific surface offering this support. This because it will represent only a part of the spatial representation in which book and computer are involved: if $g(\textbf{on'})=\textit{on}$ and $\textit{on}\leq\textit{on-top}$, then we will have $g(\textbf{on'})\leq\textit{on-top}$ to hold. In words, *on* represents only a part of a certain extra-linguistic scenario, and thus will be less accurate than *on top of*. Conversely, the relation $\textbf{on'}\leq f(\textit{on-top})$ also holds, i.e. *on* is less accurate than the adposition which would perfectly match the said scenario. Hence, the *part-of* relation, when it is defined on 'mixed' objects by means of a Galois connection, can be interpreted as relation expressing a degree of accuracy of a sentence, an adposition or any part of speech, with respect to the extra-linguistic context.

This proposal on the vision–language interface makes two main predictions. First, it predicts that the 'amount' of spatial (visual) information expressed by a sentence is flexible, and may be as accurate as the corresponding visual scenario, but also that the same scenario can be described by adpositions of different 'accuracy'. Second, it predicts that, since the 'binding' between the two layers of information may go in both directions, there is no 'causal' relation between these different computations, so one type of information is processed *independently* of the other. We are quite able to evaluate whether what we see refers to (or matches with) what we say and *vice versa*, but both mental processes need not a constant, unconscious feedback between the two levels of comprehension to occur. In words, we can say a lot about 'where' things are (including, but not limited to, geometric relations), but need not to limit ourselves to what we see.

A formal treatment of this 'parallel' processing can be represented as follows:

(29)	I	V	$V \Leftrightarrow D$	D
	$t.$	$(a+b)$		$(\mathbf{a}'+\mathbf{b}')$
	$t+1.$	$(a+b)=g(\mathbf{a}'+\mathbf{b}')$		$f(a+b)=\mathbf{a}'+\mathbf{b}'$
	$t+2.$	$(a+b)=g(\mathbf{a}'+\mathbf{b}') \Leftrightarrow f(a+b)=\mathbf{a}'+\mathbf{b}'$		

In words, at some interval in a computation, the two types of information are first mapped onto the other domain, and then (dynamically) bound together if the two 'flows' of the process yield the same result, possibly compared in terms of accuracy in a common logical space, which is represented as '$V \Leftrightarrow D$'. Informally, we check if what we see matches with what we say and *vice versa*, hence obtaining a 'broader' picture of facts. Since what we see needs not to match with what we say, the binding relation between these two types of information is entirely optional and, as we have discussed so far, it ultimately represents a top-down translation process, which can be more or less accurate.

One important thing to note is that this formal treatment is modular also because the binding of two types of information is explicitly represented as a distinct result of a matching operation. If we would have assumed that the binding occurs by the simple co-synchronous occurrence of these operations, our architecture would actually have been *connectionist*, in nature. While the two processes are isomorphic and can be tightly connected, they are nevertheless two distinct processes, and a third process is their matching relation (i.e. binding); see Marcus (2001) for discussion. Now that we have gone through the formal details and their predictions, we can focus on their empirical support, which I shall analyze in the next section.

4.2. *Testing the Theory against the Data*

The theory I have proposed in the previous section is consistent with general assumptions about vision and language as parts of a cognitive and modular architecture (cf. e.g. Jackendoff 1997, 2002), and possibly offers a more fine-grained and formally precision analysis and representation of these modules and their processes. In this section I shall explain more in detail why this theory is consistent with previous proposals and offer an 'improved' model of their insights, and why it is consistent with general assumptions about cognitive architecture, i.e. why the two main predictions I offered in the previous section hold. I shall focus on four topics, offering evidence that confirms these predictions.

A first topic pertains to the 'amount' of space found in language. Let me repeat (13) and (14) as (30) and (31) to illustrate the point:

(30) The book is on the tip of the left edge of the blue table.

(31) The book is on the table.

The crucial difference between (30) and (31) is that both sentences may be used to convey information about the same extra-linguistic scenario, but (31) is definitely more accurate than (30). Vision-wise, a scenario in which the book is supported by the tip (of the edge) of the table is also a scenario in which a book is supported by the table — hence, the relation *on≤on-top* holds. language-wise, the DRS representing (31) is part of the *DRS* representing (28), so the relation ***on'≤on-top'*** holds. Hence, the following identities *g(**on-tip'**)=on-top* and *g(**on'**)=on* hold, as well as ***on-tip'****=f(on-top)* and ***on'****=(on)*. We can then observe that the relation *g(**on'**)≤on-top* holds, i.e. that (31) is a partial representation of the same scenario that (30) is a total representation of, and thus a less accurate description of facts. Conversely, the relation ***on'****≤f(on-top)* holds, i.e. (31) expresses part of the information expressed by (30), and thus of the scenario that (30) represents.

A second topic pertains to the different degree of accuracy that two sentences can have in describing a certain scenario, when involving different adpositions. If the meaning of two adpositions overlaps or stands in an entailment relation, then speakers may favor one over another, when they need to associate it to visual information. The entailment cases are quite intuitive, and can be seen as a general case of the relation between (30) and (31). In a situation in which a book is supported by the upper part of a drawer, *on top of* may be judged as a 'perfect' adposition to describe this situation while *on*, that is entailed by *on top of*, may be considered as less appropriate, with respect to the scenario it purports to match with.

The cases in which adpositions overlap in meaning require some more discussion. Let me repeat (10) and (11) as (32) and (33) to illustrate the point:

(32) The painting is on the wall.

(33) The painting is in the wall.

In a scenario in which a panting is literally encased in the wall, (33) may be a more accurate sentence to describe this scenario than (32), because it may express in a more precise way the matching extra-linguistic scenario. Intuitively, if a painting is in the wall, it is certainly supported by it, and actually part of the wall's surface, rather than just adjacent to it (as for *on*). Formally, we can say that *in* is more accurate than *on* with respect to the aforementioned scenario if the following holds: if **in'≤*on'***, then ***on'*∩*in'*=*in'***, i.e. *in* is a part of *on* and its meaning; and thus, if *g(**on'**)≤g(**in'**)*, then *g(**on'**)∩g(**in**')=g(**in'**)*, i.e. *in* describes a more specific scenario than *on*, and is hence considered more accurate.

The treatments I discussed in the first and second topic are consistent with results like those of Coventry & Garrod (2004), Regier *et al.* (2005), and much of the aforementioned literature on spatial sentence processing, which also cover the relations between e.g., *above* and *on*, *in* and *under*, and so on. It is also consistent with Levinson & Meira's (2003) cross-linguistic results, which are indeed based on how adpositions can be conceptually organised in terms of increasing accuracy and specificity of their use in (implicit) context.[20] This literature also offers indirect evidence of the validity of my proposal: Most experiments aim to test how participants consciously match visual stimuli with linguistic stimuli, evaluating how accurate sentences can be in describing a scenario. Hence, it indirectly supports the view that the functions *f* and *g* represent conscious processes.

A third topic pertains to a complex case, that of the relation between vision and language with respect to reference systems and their computation. Again, works like Carlson-Radvansky & Irwin (1994), Carlson (1999), or Regier *et al.* (2005) show that, when speakers interpret axial terms such as *to the left of*, their accuracy can be measured with respect to different reference frames, e.g., whether a chair is to the left of a table with respect to the observer (relative frame), the chair itself (intrinsic frame), or an environmental cue like the floor (absolute frame). What I have suggested for 'standard' adpositions can be extended to these 'axial' adpositions as well, with no need to make any further assumptions. Furthermore, although some proposals conjecture that the 'cognitive' procedures by which 'absolute' spatial relations are computed dramatically differ from other visual procedures (e.g. Levinson 2003), their mapping onto linguistic unit seems to be rather 'ordinary'. Whether we may compute a polar direction such as the one corresponding to *North* via an entirely different set of cognitive resources than the ones involved in e.g., computing the support relation corresponding to *on*, the two adpositions share the same underlying grammar, and seem not to reflect this 'cognitive difference', if it exists.

From these three topics we can observe that the first prediction of my novel interface approach, the flexibility of this interface, is substantially borne out. This allows to make a further general comment regarding the "how much space" problem, and how we may choose the degree of accuracy we want to express. The literature gives us the relevant answer regarding how this process comes about, in the guise of theories of sentence planning and production. For instance, in a theory of sentence-planning (*speaking*) like Levelt (1989), speakers are assumed to decide, at a pre-linguistic level, both which basic 'facts' and the relations between these facts they wish to convey (Levelt's level of *macro-planning*), and consequently which language-specific rules (syntactic and semantic alike) to use in order to convey these facts (Levelt's level of *micro-planning*).

For our discussion macro-planning represents the relevant aspect of production, since it indirectly defines "how much" we may express about extra-linguistic information. In slightly more formal terms, macro-planning may be treated in the following way. A speaker may look at a certain general visual context *V* and may decide to express part of this scenario via the selection of a certain VRS *v*. Given a selection function *s*, this process can be represented as, for

[20] A conjecture is that classical results of prototype theory (e.g. Rosch 1975) may actually find a formally precise account, if we, for example, pursue the intuition that a noun such as *robin* may be seen as the perfect linguistic label for the sum of all visual/cognitive information we ascribe to birds, rather than *penguin*. This intuition is actually pursued in Ganter & Wille (1998) and especially in van Eijck & Zwarts (2004) in thorough detail.

example, *s(V)=v*. For instance, a speaker may look around a room and may decide to say that a certain specific book is on the tip of the left edge of the blue table. The selected VRS *v* would actually stand for the complex VRSs representing book, blue table, edges, and tips, and the relations holding between these VRSs.

The sentence corresponding to this VRS, which we can represent as *f(v)=**S′*** and thus as *f(s(V))=**S′***, indirectly represents which pre-linguistic facts are chosen by the speaker as finding their way into language. The amount of space finding its way into language roughly corresponds to the speaker's intentions to be more or less accurate in describing a scenario and his eventual desire to express one outstanding aspect over another. Although he may do so via different microplans, i.e. via the choice of different words and sentence, this choice is inherently flexible, rather than dictated by constraints on what type of spatial information finds its way in language. This is captured by the function *f* taking the function *s* as its input. Informally, we may decide to say something about the scene we are paying attention to and, in doing so, we *selectively* (and consciously) pick out visual information about this scene, then 'convert' it into the corresponding sentence, thus effectively deciding how much 'space' gets into language.

A fourth topic pertains to the relation between vision and language in case of cognitive impairment in one of the two modules. The intuition is the following: If my theory can predict how the vision–language interface works, it should also make predictions about the problems that could arise when these modules are not properly interfaced — it should be *breakdown-compatible* (e.g. Grodzinsky 1990). The following examples suggest that this is indeed the case.

A well-known fact is that people affected by Williams syndrome may have relatively intact language skills, including a good understanding of spatial language, but are usually unable to assess even basic spatial relations from a visual perspective. These patients may be able to understand an adposition such as *in front of*, but may not be able to evaluate what is the front of an object (e.g. Landau & Hoffman 2005 and references therein).

An obvious account, in the current proposal, is that, since spatial vision is quite impaired, it will not be possible to have a visual input that will correspond to a linguistic output. That is to say, the function *f(v)* will be undefined, since it will have no input, and so the function *g(**d′**)* will be undefined as well. As a consequence, it may not be possible for individuals with Williams syndrome (to give one example) to relate what they see to what they say. As it stands, my proposal seems to be consistent not only with a general modular approach to cognition, but also with a general approach to cognition and its disorders.

Another well-known case of a cognitive disorder affecting one side of the 'space' interface is aphasia. In Broca's aphasia, omission of prepositions (among other functional words) is well attested, while spatial vision is usually (completely) spared. Adposition omission in aphasia may be gradual and patients tend to omit more general adpositions (e.g. *at*) rather than less general adpositions (e.g. *in front of*; see e.g. Trofimova 2009 for a recent review). Regardless of their degree of language impairment, aphasics usually lose their ability to produce but usually not their ability to comprehend adpositions, and, more generally, language; hence, they are able to understand whether adpositions correctly describe a scenario or not. While one aspect of spatial language can be dramatically impaired (e.g. production), all other aspects of both spatial vision and language, including their interface, are substantially spared, in line with my assumptions.

A similar account may be extended to another cognitive disorder, that of dyslexia.[21] Models like the *Dual Route Cascaded* model of reading aloud (*DRC*; e.g. Coltheart *et al.* 2001 but also Beaton 2004), the processing of ('reading') a single word is assumed to occur via three parallel processes, one in which we visually recognize a written word (*non-lexical route*), and one in which we (may) retrieve its lexical entry as well as its phonological and syntactc-semantic properties (*lexical/sub-lexical route*). Although one process can be faster than the other, full recognition of a word occurs when both processes converge to the same output, but fails if the 'visual' process is damaged (failure to read graphemes and words, or *shallow dyslexia*) or the 'linguistic' process is damaged (failure to understand the meaning of words, or *deep dyslexia*).

As per the other cognitive disorders, our theory of the vision–language interface is consistent with this analysis of dyslexia without any further assumptions. Although for dyslexia we would certainly need a more accurate and specific analysis of both sides of the problem, the intuition seems to be correct: We may not be able to see certain visual objects correctly, but we may still retrieve their corresponding linguistic labels, and *vice versa*. We can also observe that the second prediction is borne out, since these cognitive disorders show that spatial computations can occur both at the visual and linguistic level and can be bound together, but also that this binding process is not necessary. In fact, even if one side of this process may be completely impaired, the other side will be still able to work independently.

Summing up, the discussion of these four topics suggests that our vision–language interface theory can have theoretical value and can withstand empirical scrutiny, even once we look beyond the topic of space. As we have seen, visual and linguistic representations can be matched in a quite precise way, but the processes regulating this matching of information is inherently conscious, that is, based on a speaker's top-down thought processes. Speakers may wish to be more or less accurate in describing a scenario and may evaluate sentences with respect to their descriptive accuracy. They may be able to understand spatial language even if they can't navigate the environment and, for complex tasks such as reading (i.e. the codified matching of visual and linguistic stimuli), they require conscious and protracted effort to establish the proper mappings, provided that this mapping is not impaired by cognitive deficits.

These facts are somehow hard to explain in previous accounts of the vision–language interface but fall out as predictions of the theory I have sketched so far due to its flexibility. This theory also presents in detail the convergences between space, vision, and language, offering a view in which these two modules are remarkably similar; as such, it may appear that there is little or no difference between the two modules, both from a structural and a content-bound point of

[21] Dyslexia can be informally defined as a cognitive disorder which influences our ability to successfully read, i.e. to either successfully decode the sequence of graphemes ('letters') making up a written word, or to properly interpret a word, and access syntactic information about it. See Beaton (2004) for a thorough introduction.

view. I shall focus on these differences in the next section.

4.3. *What Is Unique to Language, and Why*

The discussion I have offered so far has sketched the strong similarities between vision and language as modules of cognition. It has also offered an attempt to explain how these two modules exchange information — for instance, via the synchronization of their processes. (Spatial) vision and language seem to be remarkably similar modules, and it is not surprising that in some quarters they are considered as contiguous modules, if not the same module, in some respect (e.g. Talmy 2000, Coventry & Garrod 2004).

There are, however, a number of properties of language which seem rather hard to reduce to general, non-linguistic features, and which inherently involve the possibility in language to convey information about 'unbounded' quantities. In the common parlance of biolinguistic research, much of our discussion up to this point has focused on defining the properties that can be ascribed to the broad faculty of language (FLB), in the terminology of Hauser *et al.* (2002), since I have mostly been concerned with the relation between vision and language, and with those properties that are shared by both computational systems. In this section, I shall sketch a very preliminary proposal, stemming from the discussion offered so far, on what properties are unique to language and thus may be possible candidates to form the kernel of the faculty of language in the narrow sense (FLN). I shall do so by focusing, for the most part, on spatial language. I shall discuss these properties in a less formally rigorous way, focusing on speculative aspects of the discussion.

Look at the examples:

(34) Mario has gone to the store *three times*.

(35) Mario *may* go to the store.

(36) *All* the boys have gone towards the store.

(37) *Every* boy will go toward the fence.

(38) *A* boy may come to the party.

(39) *Some* boy may come to the party.

(40) Mario *always* goes to the store.

(41) Mario *seldom* goes to the store.

(42) *Where* are you going?

(43) I am going *there*, too.

(44) Mario *lends* a book to Luigi.

(45) Luigi *borrows* a book from Mario.

In (34), Mario's going to the store is described as occurring three times or instances, but little is said about when this happens: It may occur one time right now, one time yesterday, and one time when he was a young lad. Two of the events that the adverb denotes cannot be mapped onto visual inputs, because two of them cannot correspond to current facts, but rather to 'memory traces' we have recorded of them. Language allows us to merge together pieces of information which do not necessarily correspond to one modality, into a unified type of information.

In (35), Mario's possible event of going to the store is something that we conceive as occurring in, say, a few more minutes, or whenever he feels like it, perhaps tomorrow. In the case of the non-current events of (34), the modal auxiliary may simply denote a linguistic unit which hardly can find a visual unit as its counterpart. In (36) and (37), the amount of boys that have gone to the store may vary, and may involve pairs or triples (or bigger quantities), but each of these possible combinations of boys will go to the store, without any exceptions.

In (38) and (39), instead, we may not know the identity of who is going to come to the party, except that it is likely to be a single boy, someone who we may have not mentioned so far and may never come to know, let alone see. These cases may already show that the mapping from vision to language can be quite partial (i.e. not always defined), but the following cases should give even stronger evidence. Adverbs such as *always* and *seldom*, as in (40) and (41), suggest that we may even convey linguistic information about several (infinite) situations (sets of events) in which Mario goes to the store, or say that such situations are rare but do occur (i.e. *seldom*).

Examples like (42) and (43) show that we may actually rely on someone else's ability to access information in order to retrieve information of Mario's whereabouts: If someone answers our question, we will be able to know Mario's location without actually seeing this location, and if someone has already told us *where* Mario is going, we may say that we are going *there*, although we may not be able to see "where" "there" is. In (44) and (45), the same set of events is presented under two different, and in a sense complementary, perspectives: While the visual scenario is in a sense the same (a book is temporarily exchanged, between Mario and Luigi), the two sentences express these facts from Mario or Luigi's perspective, respectively.

There are two generalizations that we can make from these examples. One is that language may convey information which can be *multi-modal*, in the sense that linguistic units may bring and represent together information which comes from different cognitive sources, and may have no extra-linguistic instantiation whatsoever. This is not surprising if we look at language at a module that only processes internal information, stripped of any perceptual or modal-specific aspects (unlike vision), but it is also consistent with various theories of memory as a 'mental' model in which we record and organize memory.

One way to look at this aspect is the following, and it is based on theories of memory like Cowan's (1988, 1995, 2005). In this theory, long-term memory is seen as the model representing all the information we may have stored about the world, whether it is veridical or not (i.e. whether it is represented in episodic memory or not[22]). Short-term memory, on the other hand, can be seen as the current

[22] Episodic memory is a component of memory which 'records' perceptual information regard-

part of long-term memory which is accessed and evaluated at a given time. In our logic, long-term memory can be seen as a static model *<D>* or *<V>*, while short-term memory can be seen as the dynamic counterparts of these models, *<I,D>* or *<I,V>*.

For instance, we may have observed Mario going to the store in three very different moments of our life, but if we use a sentence like (32), we represent these otherwise separate events of time in the same representation (ultimately, a DRS) in our short-term memory. Language allows us to define a 'common space' in which 'displaced' events form a may form a consistent representation insofar as they share the same formal properties (e.g., being three instances of a walking event), and thus are stripped of any constraints on perceptual information — but may also be bound with other 'portions' of short-term memory (e.g., visual computations; cf. the previous section). Informally, an adverb like *three times* says that there are three contiguous intervals in a derivation in which three events of going to the station become logically contiguous, i.e. we have *a+b+c* at an interval *t+n*.

Another generalization is that language can express relations and quantities which are not necessarily finite (or bounded), and is not limited to offering one perspective. This latter, (quite) rough, intuition is based on our last pair of examples, but several other similar examples could be made; think of any active sentence and its passive counterpart, for instance. If we think in slightly more formal terms, we may construe (42) as representing a scenario in which Mario's actions as an 'agent' operates onto Luigi as a 'patient', and can be schematically represented as $a \rightarrow p$. We can then assume that (43) can be represented as the inverse type of relation, which can be represented as also $\neg(a \rightarrow p)$. In very informal words, we can represent that the sequence of events expressed by (43) flows in the opposite direction of (42), as the informal use of negation aims to represent, although we express the order of relevant entities in the same way as in (42).

This is possible because in language we can express the same underlying conceptual structures under different 'perspectives', but via virtually the same logical apparatus (cf. also Landman 1991: chap. 3 and 2004: chap. 7–8). Again, if we think of language as defining a conceptual 'space' not constrained by perceptual limits, then the same underlying information can be expressed in two apparently opposite ways, which, however, underlie the same logical principles (e.g., *monotonicity*). Although (42) and (43) describe the same extra-linguistic event, their interpretations represent two possible ways by which language can structure this information.

Another form of 'unboundedness', its linguistic realization as well, is ultimately represented by the interpretation of quantifiers and other 'expressions of quantity', as the examples show. Informally, in language we can express information about numbers of individuals which are far greater than the amount of individuals we can 'see' and which can be structured in rather complex and fine-grained ways, as adverbs like *seldom* and *always* suggest.

This can be illustrated via a detailed analysis of (34) and (35). Note here that I shall depart quite dramatically from DRT and treat a quantifier like *every* as represented by a logical operator rather than a duplex condition. In both senten-

ing the first time we observe a given event.

ces, it is possible to represent the contribution of *all* and *every* to the sentence in terms of the universal quantifier, which I shall here represent in its Boolean incarnation, '∧' (e.g., Montague 1973, Keenan & Faltz 1985). This symbol can be informally interpreted as a form of unbounded coordination: Informally, the sentence *Every boy has gone to the store* can be interpreted as the equivalent of "Mario has gone towards the store and Luigi has gone to the store a*nd…*", i.e. as if we were to state each possible boy in a large, perhaps infinite, domain of discourse, one by one.

Suppose then that we take the set of boys as a list (sequence) of boys in discourse. The DRS representing *all the boys* is equivalent to the merging of the DRS representing the sum of the last boy with the sequence of boys occurring before him in this infinite list. We define the interpretation of a universally quantified noun phrase (its DRS) via the sum of the interpretation of its parts, via induction (its constituting DRSs).

This can be represented as:

(46) *t.* [{x–2}:**boy′**(x–2)]+[{x–1}:**boy′**(x–1)]=
t+n. [{x–2,x–1}:**boy′**(x–2),**boy′**(x–1)]=
t+n+1. [∧x:**boy′**(x)]

With the referent/individual *(x–2)* representing the list of boys preceding the last boy (i.e. the second-to last (complex) referent), *(x–1)* representing the last boy, and ∧*x* representing the 'new' referent obtained from the merging of the two 'old' referents. This is a *recursive, inductive* definition of the universal quantifier, in terms of an unbounded form of *merge*, and its interpretation. In words, we interpret the universal quantifier as the result of taking each referent in discourse via one common condition. This result is another DRS, the DRS representing the result of taking each referent which can be identified as a "boy" one by one, i.e. via the product of each condition merged in a DRS, here represented as '∧'.

These considerations can be also extended to other quantifiers with the proper *provisos*, but also to adpositions, thus suggesting that spatial language is also 'unbounded' in its interpretive range. For instance, the relational component of any adposition (e.g. *near*) can be recursively defined as the merging of two opportune relations. Abstracting away from the specific condition on proximity (i.e. **near′**) and with some notational fantasy, "near" can be represented as:

(47) [{s,s′}:s≤s′]=[{s,s′–2}:s≤(s′–2)]+[{s,s–1}:s≤(s′–1)]

Here the 'geometry' approach to adpositions is quite useful to illustrate the intuitive meaning of (47). If a figure is the ground when it occupies a certain region, then it will be near the ground if it occupies any region which is included in the bigger region. Conversely, once we sum all the (sub)-regions in which a figure is near a ground, then we will obtain the 'general' region which can be labeled as *near*.

This way of representing the universal quantifier, and in general of representing quantified noun phrases, as well as the interpretation of *near* and other adpositions, is informally based on one recursive function, the *Fibonacci series*, which allows to define one object (e.g., a natural number) as the sum of its direct predecessors. Intuitively, it may be extended to all of the other functional words I have discussed in examples (34)–(45), and to any expression that captures a form of quantification.

Several authors have argued that the Fibonacci series can represent how recursion is expressed in language (e.g. Soschen 2008 and references therein), but one may also assume that the successor function may be a recursive function that can also be used to represent the recursive nature of syntactic processes (see Landman 1991: chap. 1 for discussion). The crucial aspect is that, since language is different from vision by being fully recursive at a syntactic level, it will also be different in having terms which directly express the result (interpretation) of this unboundedness, and thus will be fully recursive at the semantic level.

An indirect way of capturing this difference is by enriching our logic representing language, so that we have the tuple $L{=}{<}D,+,\leq,\wedge{>}$. This tuple represents the 'structure' of language as including not only a minimal syntax (*merge*, '+') and semantics (the *part-of* relation, '≤'), but also a set of operators, here represented by the universal quantifier, that denote *the result* of linguistic processes. We have an indirect reconstruction of the distinction between FLN and FLB.

This reconstruction is indirect only because recursion is a resulting property of the 'logic of language', but it nevertheless represents one (maybe *the*) element of distinction between vision and language. Informally, it tells us that language has certain recursive *closure principles* which allow to label not only objects from other models (e.g. nouns for objects), but also to express the processes by which we collect together these objects into abstract structures. Adpositions represent one case, and quantifiers represent a more language-specific case, but the same reasoning could be applied to any functional word in language. We are able to talk about, for example, *all the past boys and apples* because we are able to compute a referent that stands for the combination of two different sets of entities (i.e. boys and apples), possibly representing entities 'displaced' in time and space — even if these sets may include an infinite amount of 'smaller' referents (i.e. each single boy and apple).

It is also indirect because, as we have seen, visual and linguistic information are processed as distinct types of information, although they are potentially connected up to isomorphism. While there can be an intimate relation between what we see and what we say, language is not bound by other modules of cognition in its expressive power, although the entities that make up the universe of discourse denoted by language must be the result of previous processes of interpretation, as the closure principle entails.

One important aspect, however, is again that vision can be represented via a similar, although less 'powerful', logical structure: As observed in Pinker & Jackendoff (2005) and Jackendoff & Pinker (2005), a number of 'structural' or hierarchical properties are *domain-general*, and thus not unique to language, because they represent domain-general logical principles by which we process, retain, and organize different types of information. Vision represents here one important case, but the phonological component of language also offers a similar case, and other examples abound (e.g., the 'grammar of action' analyzed by Fujita

2009, the ‘grammar of music’ in Jackendoff & Lerdahl 2006, or the ‘grammar of phonetics’ of Reiss 2007).

The intuition behind these considerations is the following. Each module of cognition that is properly definable can be represented via the same underlying logic, which I have presented here in two slightly different ‘incarnations’. The structures defined by this logic are models of the ‘things’ they represent, for instance visual objects. These models can be infinite, since they can potentially represent, for example, the infinity of objects we can recognize, events we can witness, and so on. The models defined by each module can be mapped onto a ‘common’ logical space, that of language: We can talk about what we see, smell, think, believe, etc.

This very informal discussion can be made more precise via the discussion of a well-known theorem of model-theoretic semantics, the *Löwenheim–Skolem theorem*. This theorem can be very roughly paraphrased in the following way: If a first-order logic has infinite models, then it is a corresponding countable infinite model. In our case, its import can be seen as follows. We may define several logical systems, each of them representing a single module of cognition. Each logic has the same underlying (and thus domain-general) syntactic and semantic principles. Each logic can define an infinite model: We may be able to recognize an infinity of (moving) objects, an infinity of sounds, realize an infinite of possible actions, and so on. Defined in this way, each logic/module appears to be an independent system, an internal model that potentially allows to represent how we can interact with the external world, but needs not to rely on ‘external’ inputs for these computations.

The Löwenheim–Skolem theorem tells us that even if we have an infinity of such logical systems, it is possible to define a more general logic which includes all of these modules in a ‘common’ logical space. More precisely, the *downward* part of the theorem tells us that, if a model is (countably) infinite, then this model may include an infinity of possible sub-models, themselves infinite. The *upward* part of the theorem tells us that for each (infinite) sub-model, we can find an extension of this model that includes the sub-model and some other elementary statements. So, if our ‘main’ model represents language, it will include models of other modules as proper sub-models (downward part); if a module such as vision can be represented via a model, then this model can be integrated inside the main model of language (upward part).

The conceptual import of this theorem can be dynamically interpreted as follows. We can assume that, for each (well-formed) visual computation, we can have a matching VRS in our model of vision. Each visual information unit can then be mapped onto the language model, and thus be part of a general model that includes other types of information (upward part). Conversely, for each linguistic unit so defined, a corresponding non-linguistic unit can be found, so that from the general model, we can move to the more specific model (downward part). This process can unravel over time: for each thing we see, we may have a corresponding noun, which we then associate to any object that has that shape, to put it in a very informal way. The same principle of closure can be defined for adpositions (and verbs): For each type of spatio-temporal relation between objects we ‘see’, we may have a corresponding adposition, which we associate to any relation that has that spatio-temporal structure, or 'shape'.

Both model (language) and sub-model (vision) will thus be expanded or updated over time, but the underlying (Boolean) structure representing these processes and their results will retain the same basic structure, as this update process will be guided by the same basic principles. Informally, these models can become quite 'rich' over time, but the basic *structural principles* by which their growth occurs remain the same, as a consequence of their recursive definition. In this regard, (full) recursion represents the possibility for the language to apparently expand *ad infinitum*, representing any type of information in a common space. Similarly, the relation between the language model and its sub-models, which takes the shape of interface relations/conditions, represents the possibility that language (recursively) emerges as a 'general' model, generated by the projection of all models of cognition into a 'neutral' logical space.[23]

I shall thus propose the following answer to the third research question:

A–3: *The nature of the vision–language interface is that of a bijection; recursive closure principles and interface conditions define what is unique to language.*

What distinguishes language from other modules of cognition is not the type of underlying structure, but two properties emerging from this structure and its ability to represent other structures in common space, (full) recursion and interface conditions. The answer I offered so far is virtually the same offered by Hauser *et al*. (2002), although the argument on which I have based my answer is relatively different, and perhaps places a greater emphasis on the *interaction* between recursion and interface conditions and their inherent 'logicality', as the kernel properties of FLN. This answer is also consistent with the considerations made by Pinker & Jackendoff (2005) and similarly-minded contributions to the FLN/FLB debate, since it suggests that language and other modules of cognition are quite more similar than it may appear at first glance.

The answer I offered so far might also offer an insight with respect to one important Biolinguistic problem, the emergence of language from an evolutionary perspective. I tentatively suggest the following way to look at this problem, assuming in advance that what I shall say in this paragraph is nothing more than a wild (and perhaps wrong) conjecture. If we take a logical perspective and compress the evolutionary millennia into a conceptual space, then the emergence of a FLN kernel, from an evolutionary perspective, occurs when the integration of different types of information into a common format emerges.

Pushing this speculative line to its logical limit, we might assume that at some point, roughly 200,000 years ago, our ancestors were (suddenly?) able to compare what they saw with what they said, and *vice versa*. From this initial step, which could be called the first step of induction, we might as well as assume that the $n+1^{\text{th}}$ subsequent steps followed suit over the next few hundred of thousand years, taking shape as the unraveling of the gamut of languages we can currently

[23] This assumption leaves open the problem of 'how many' models make up our cognitive architecture that are integrated in this model. I leave open this question, but I assume that we can leave out the 'massive modularity' hypothesis typical of some evolutionary psychology literature. See Fodor (1998, 2000) for further discussion, however.

attest in the world (see e.g. Piattelli-Palmarini & Uriagereka 2005 for discussion and some references).

This single and yet very powerful emergent property could have arisen as the possibility (perhaps, necessity) to integrate different bits of information into an internally coherent (and perhaps optimal) representational/computational system. It might have been the case that language arose as the 'proof' that it is possible for different cognitive processes/modules to combine together into a unified, coherent cognitive architecture; thus it emerged entirely because of internal, structural pressures (Buzśaki 2006), although it became one tool (out of many) for humans to grasp and represent facts about the world, including the position of the things we see around us.

I shall leave these complex topics aside, and focus my attention back to our much more modest topic of discussion. Given the discussion I offered so far, I shall propose the following answer to the global research question:

Q-A: *The relation between spatial vision and spatial language is an isomorphism, as both models represent the same 'amount' of information via different types of information.*

This answer sums up the results of this section. Note that, while in this section I have suggested that language, broadly defined, describes a model which includes a model of vision as one of its proper parts, if we focus on spatial language, then this portion of language has the same structure and properties of vision; consequently, it (correctly) appears that vision and language are more similar than it seems, as observed in much literature. Much more could be said about this topic, as the discussion I have offered in this section can only be thought as a very preliminary attempt at refining a notion of FLN (and FLB, for that matter) and its emergence, from the point of view of 'space'. For the moment, though, I shall leave this discussion aside, and move to the conclusions.

5. Conclusions

In this paper, I have offered a novel proposal regarding the relation between vision and language with respect to 'space' — our understanding of things and their place in the world. I have argued that our spatial vision and language are quite abstract in nature, as they involve the processing of various types of information and their ability to individuate objects and the events they are involved in as well as the 'structural' relations that emerge from this process of individuation.

In doing so, I have offered a number of innovations on several closely related topics, including an updated review of the debate, a model-theoretic approach to vision which covers data usually ignored in the debate on 'space' (via the VRT proposal), a novel DRT treatment of adpositions, and a novel analysis of the vision–language interface, and what consequences this analysis has for a general theory of the language faculty.

The general picture I offered is one in which different models of cognitive processes can be formally defined in detail, and then embedded into a more general model of 'knowledge', modeled via a particular approach to Fodor's (1975) notion of 'language of thought', DRT, and the 'modularity of mind' hypothesis (Fodor 1983), although taken from a definitely more logical stance (as in e.g. Crain & Khlentzos 2008, 2009). Informally, language represents a 'neutral' logical space, a model of knowledge representation in which different concepts can be freely combined together, since they are already stripped of their 'external' constraints when they are represented in the corresponding models (e.g., Asher & Pustejovsky 2004, Asher 2011, and references therein). A similar reasoning holds for the articulatory-perceptual side of language. While we need to organize speech streams, say, into coherent units, the result of this process must then be organized into a coherent structure of syllables, words, and utterances which may be organized according to processes and relations not unlike those of other modules, and which are then mapped onto concepts, and thus lose their 'external' part. See Reiss (2007), Hale & Reiss (2008), and Samuels (2009) for discussion.

In this regard, language is the model that comes into being when all other 'sub-models' expressed by other modules of cognition are joined in a common logical space, and which might have emerged as the 'projection' of different cognitive modules into this common logical space. With respect to this neutral logical space, then, spatial language represents that fragment of space which represents spatial vision, i.e. our abstract representation of things in the world, whether this representation is veridical or not. As a consequence, the proposals I have made here, although still very preliminary in their nature, can be seen as offering a better picture not only on what is the nature of spatial representations in vision and language, but also on the logic behind the processes by which we combine together these representations, and what this tells us about the general architecture of mind and language.

References

Arsenijević, Boban. 2008. From spatial cognition to language. *Biolinguistics* 2, 23–45.

Asher, Nicholas. 2011. *Lexical Meaning in Context: A Web of Word*s. Cambridge: Cambridge University Press.

Asher, Nicholas & James Pustejovsky. 2004. Word meaning and commonsense metaphysics. Ms., University of Texas at Austin.

Barwise, Jon & John Etchemendy. 1990. Information, infons and inference. In Robin Cooper Kuniaki Mukai & John Perry (eds.), *Situation Semantics and Its Applications*, vol. 1, 33–78. Stanford, CA: CSLI.

Barwise, Jon & Jerry Seligman. 1997. *Information Flow: The Logic of Distributed Systems*. Cambridge: Cambridge University Press.

Beaton, Alan. 2004. *Dyslexia, Reading and the Brain: A Sourcebook of Psychological Biological Research*. Hove: Psychology Press.

Biederman, Irving. 1987. Recognition by components; a theory of human image understanding. *Psychological Review* 94, 115–147.

Bierwisch, Manfred. 1996. How space gets into language? In Bloom *et al.* (eds.), 1–30.

Blackburn, Paul, Johan van Benthem & Frank Wolter (eds.). 2006. *Handbook of Modal Logic*. Amsterdam: Elsevier.

Bloom, Paul. 2000. *How Children Learn the Meanings of Words*. Cambridge, MA: MIT Press.

Bloom, Paul, Mary A. Peterson, Lynn Nadel & Merrill Garrett (eds.). 1996. *Language and Space*. Cambridge, MA: MIT Press.

Bruce, Vicki, Patrick R. Green & Mark A. Georgeson. 2004. *Visual Perception: Physiology, Psychology and Ecology*, 4th edn. Hove: Psychology Press.

Burgess, Neil. 2006a. Spatial memory: How egocentric and allocentric combine. *Trends in Cognitive Science* 24, 603–606.

Burgess, Neil. 2006b. Spatial cognition and the brain. *Acts of the New York Academy of Science* 1124, 77–97.

Burgess Neil, Eleanor A. Maguire & John O'Keefe. 2002. The human hippocampus and spatial and episodic memory. *Neuron* 35, 625–641.

Burgess, Neil & John O'Keefe. 1996. Neuronal computations underlying the firing of place cells and their role in navigation. *Hippocampus* 6, 749–762.

Burgess, Neil & John O'Keefe. 2003. Neural representations in human spatial memory. *Trends in Cognitive Sciences* 7, 517–519.

Buzśaki, György. 2006. *Rhythms of the Brain*. Oxford: Oxford University Press.

Carey, Susan. 1992. Becoming a face expert. *Philosophical Transactions of the Royal Society of London* 335, 95–103.

Carey, Susan. 1994. Does learning a language require conceptual change? *Lingua* 92, 143–167.

Carey, Susan. 2001. Whorf versus continuity theorists: Bringing data to bear on the debate. In Melissa Bowerman & Stephen C. Levinson (eds.), *Language Acquisition and Conceptual Development*, 185–214. Cambridge: Cambridge University Press.

Carey, Susan & Fei Xu. 2001. Infants knowledge of objects: Beyond object files and object tracking. *Cognition* 80, 179–213.

Carlson, Laura A. 1999. Selecting a reference frame. *Spatial Cognition and Computation* 1, 365–379.

Carlson, Laura A., Terry Regier & Eric Covey. 2003. Defining spatial relations: Reconciling axis and vector representations. In van der Zee & Slack (eds.), 111–131.

Carlson, Laura A., Terry Regier, William Lopez & Bryce Corrigan. 2006. Attention unites form and function in spatial language. *Spatial Cognition and Computation* 6, 295–308.

Carlson, Laura A. & Emile van der Zee (eds.). 2005. *Functional Features in Language and Space: Insights from Perception, Categorization and Development*. Oxford: Oxford University Press.

Carlson-Radvansky, Laura A. & David E. Irwin. 1994. Reference frame activation during spatial term assignment. *Journal of Memory and Language* 33, 646–671.

Chierchia, Gennaro. 1995. *Dynamics of Meaning: Anaphora, Presupposition and the*

Theory of Grammar. Chicago: Chicago University Press.

Chierchia, Gennaro & Raymond Turner. 1988. Semantics and property theory. *Linguistics & Philosophy* 11, 191–216.

Coltheart, Max, Kathleen Rastle, Conrad Perry, Robyn Langdon, & Johannes Ziegler. 2001. DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, 108, 204–256.

Coventry, Kenny R. & Simon C. Garrod. 2004. *Saying, Seeing and Acting: The Psychological Semantics of Spatial Prepositions*. Hove: Psychology Press.

Coventry, Kenny R. & Simon C. Garrod. 2005. Towards a classification of extra geometric influences on the comprehension of spatial prepositions. In Carlson & van der Zee (eds.), 149–162.

Coventry, Kenny R., Thora Tenbrink & John Bateman (eds.). 2009. *Spatial Language and Dialogue*. Oxford: Oxford University Press.

Cowan, Nelson. 1988. Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information processing system. *Psychological Bulletin* 104, 163–191.

Cowan, Nelson. 1995. *Attention and Memory: An Integrated Framework*. New York: Oxford University Press.

Cowan, Nelson. 2005. *Working Memory Capacity*. Hove: Psychology Press.

Craik, Kenneth. 1943. *The Nature of Explanation*. Cambridge: Cambridge University Press.

Crain, Stephen & Drew Khlentzos. 2008. Is logic innate? *Biolinguistics* 2, 24–56.

Crain, Stephen & Drew Khlentzos. 2009. The logic instinct. *Mind and Language* 25, 30–65.

Crain, Stephen & Mark Steedman. 1985. On not being led up the garden path: The use of context by the psychological parser. In David R. Dowty, Lauri Karttunen & Arnold M. Zwicky (eds.), *Natural Language Parsing: Psychological, Computational and Theoretical Perspectives,* 320–358. Cambridge: Cambridge University Press.

Cresswell, Maxwell J. 1985. *Structured Meanings*. Cambridge, MA: MIT Press.

Davidson, Donald. 1967. The logical form of action sentences. In Nicholas Rescher (ed.), *The Logic of Decision and Action*, 81–95. Pittsburgh, PA: University of Pittsburgh Press.

van der Does, Jaap M. & Michiel van Lambalgen. 2000. A logic of vision. *Linguistics & Philosophy* 23, 1–92.

van Eijck, Jan & Joost Zwarts. 2004. Formal concept analysis and prototypes. *Proceedings of the Workshop on the Potential of Cognitive Semantics for Ontologies*, 1–7.

van Eijck, Jan & Hans Kamp. 1997. Representing discourse in context. In Johan van Benthem & Alice ter Meulen (eds.), *Handbook of Logic and Language*, 179–237. Amsterdam: Elsevier.

Farah, Martha J. 2004. *Visual Agnosia*, 2nd edn. Cambridge, MA: MIT Press.

Fitch, W. Tecumseh, Marc D. Hauser & Noam Chomsky. 2005. The evolution of the language faculty: Clarifications and implications. *Cognition* 97, 179–210.

Fodor, Jerry A. 1975. *The Language of Thought*. Cambridge, MA: Harvard University Press.

Fodor, Jerry A. 1983. *Modularity of Mind*. Cambridge, MA: The MIT Press.

Fodor, Jerry A. 1998. *Concepts: Where Cognitive Science Went Wrong*. New York: Oxford University Press.

Fodor, Jerry A. 2000. *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology.* Cambridge, MA: The MIT Press.

Fodor, Jerry A. 2003. *Hume Variations*. New York: Oxford University Press.

Fong, Vivienne. 1997. The order of things: What directional locatives denote. Stanford, CA: Stanford University Ph.D. dissertation.

Fujita, Koji. 2009. A prospect for evolutionary adequacy: Merge and the evolution and development of human language. *Biolinguistics* 3, 128–153.

Ganter, Bernhard & Rudolf Wille. 1998. *Formal Concept Analysis: Mathematical Foundations*. Berlin: Springer.

Gärdenfors, Peter. 2000. *Conceptual Spaces: The Geometry of Thought*. Cambridge, MA: MIT Press.

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

Gibson, James J. 1966. *The Senses Considered as Perceptual Systems*. Boston, MA: Houghton Mifflin.

Gibson, James J. 1979. *The Ecological Approach to Visual Perception*. Boston, MA: Houghton Mifflin.

Grätzer, George. 1978. *Universal Algebra,* 2nd edn. New York: Springer.

Grodzinsky, Yosef. 1990. *Theoretical Perspectives on Language Deficits*. Cambridge, MA: MIT Press.

Hale, Mark & Charles Reiss. 2008. *The Phonological Enterprise*. Oxford: Oxford University Press.

Hamm, Fritz & Michiel van Lambalgen. 2005. *The Proper Treatment of Events*. Cambridge, MA: MIT Press.

Hauser, Marc D., Noam Chomsky & William Tecumseh Fitch. 2002. The faculty of language: What is it, who has it, and how did it evolve? *Science* 298, 1569–1579.

Heim, Irene. 1982. The semantics of definite and indefinite noun phrases. Amherst, MA: University of Massachusetts Ph.D. Dissertation.

Herskovits, Annette. 1986. *Language and Spatial Cognition: An Interdisciplinary Study of the Prepositions in English*. Cambridge: Cambridge University Press.

Higginbotham, James. 2009. Two interfaces. In Massimo Piattelli-Palmarini, Juan Uriagereka & Pello Salaburu (eds.), *Of Minds and Language: A Dialogue with Noam Chomsky in the Basque Country*, 142–154. Oxford: Oxford University Press.

von Hofsten, Claes, Qi Feng & Elizabeth S. Spelke. 2000. Object representation and predictive action in infancy. *Developmental Science* 3, 193–205.

von Hofsten, Claes, Peter Vishton, Elizabeth S. Spelke, Qi Feng & Kerstin Rosander. 1998. Predictive action in infancy: Tracking and reaching for moving objects. *Cognition* 67, 255–285.

Hughes, Georges E. & Maxwell J. Cresswell. 1996. *A New Introduction to Modal Logic*. New York: Routledge.

Hummel, John E. & Irving Biederman. 1992. Dynamic binding in a neural network for shape recognition. *Psychological Review* 99, 480–517.

Hummel, John E. & Brian J. Stankiewicz. 1996. Categorical relations in shape perception. *Spatial Vision* 10, 201–236.

Hummel, John E. & Brian J. Stankiewicz. 1998. Two roles for attention in shape perception: A structural description model of visual scrutiny. *Visual Cognition* 5, 49–79.

Jackendoff, Ray. 1983. *Semantics and Cognition*. Cambridge, MA: MIT Press.

Jackendoff, Ray. 1990. *Semantic Structures*. Cambridge, MA: MIT Press.

Jackendoff, Ray. 1991. Parts and boundaries. *Cognition* 41, 9–45.

Jackendoff, Ray 1997. *The Architecture of the Language Faculty*. Cambridge, MA: MIT Press.

Jackendoff, Ray. 2002. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford: Oxford University Press.

Jackendoff, Ray & Steven Pinker. 2005 The nature of the language faculty and its implications for evolution of language (Reply to Fitch, Hauser, & Chomsky). *cognition* 97, 211–225.

Jackendoff, Ray & Fred Lerdahl. 2006. The human music capacity: What is it and what's special about it? *cognition* 100, 33–72.

Johnson-Laird, Philip N. 1983. *Mental Models*. Cambridge, MA: Harvard University Press.

Johnson-Laird, Philip N. 1992. *Human and Machine Thinking*. Hillsdale, NJ: Lawrence Erlbaum.

Kahneman, Daniel, Anne Treisman & Brian J. Gibbs. 1992. The reviewing of object files: Object-specific integration of information. *Cognitive Psychology* 24, 175–219.

Kamp, Hans. 1979a. Events, instants and temporal reference. In Rainer Bäuerle, Urs Egli & Arnim von Stechow (eds.), *Semantics from Different Points of View*, 376–417. Berlin: Springer.

Kamp, Hans. 1979b. Some remarks on the logic of change. Part I. In Christian Rohrer (ed.), *Time, Tense and Quantifiers*, 103–114. Tübingen: Niemeyer.

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: Mathematisch Centrum.

Kamp, Hans & Uwe Reyle. 1993. *From Discourse to Logic*. Dordrecht: Kluwer.

Kamp, Hans, Josef van Genabith & Uwe Reyle. 2005. Discourse Representation Theory. In Dov M. Gabbay & Franz Guenthner (eds.), *Handbook of Philosophical Logic*, 125–394. Dordrecht: Kluwer.

Keenan, Edward L. & Leonard M. Faltz. (1985). *Boolean Semantics for Natural Language*. Dordrecht: Reidel.

Keil, Frank C. 1989. *Concepts, Kinds and Cognitive Development*. Cambridge, MA: MIT Press.

Kim, In Kyeong & Elizabeth S. Spelke. 1992. Infants' sensitivity to effects of gravity on visible object motion. *Journal of Experimental Psychology: Human Perception and Performance* 18, 385–393.

Kim, In Kyeong & Elizabeth S. Spelke. 1999. Perception and understanding of effects of gravity and inertia on object motion. *Developmental Science* 2, 339–362.

Kracht, Marcus. 2002. On the semantics of locatives. *Linguistics & Philosophy* 25, 157–232.

Kratzer, Angelika. 1989. An investigation of the lumps of thought. *Linguistics &*

Philosophy 12, 607–653.

Kratzer, Angelika. 2007. Situations in natural language semantics. In Edward N. Zalta (ed.), *The Stanford Online Encyclopedia of Philosophy*, Fall 2011 edn. [http://plato.stanford.edu/entries/situations-semantics].

Kratzer, Angelika. To appear. *The Event Argument and the Semantics of Verbs*. Cambridge, MA: the MIT Press.

Krifka, Manfred. 1998. The origins of telicity. In Susan Rothstein (ed.), *Events and Grammar*, 197–235. Dordrecht: Kluwer.

Landau, Barbara. 1994. Object shape, object name, and object kind: Representation and development. *The Psychology of Learning and Motivation* 31, 253–304.

Landau, Barbara. 2002. Early experience and cognitive organization. In Lynn Nadel (ed.), *Encyclopedia of Cognitive Science*. Oxford: Wiley-Blackwell.

Landau, Barbara & James E. Hoffman. 2005 Parallels between spatial cognition and spatial language: Evidence from Williams syndrome. *Journal of Memory and Language* 53, 163–185.

Landau, Barbara & Ray Jackendoff. 1993. What and where in spatial language and spatial cognition. *Behavioral and Brain Sciences* 16, 217–265.

Landau, Barbara, Linda B. Smith & Susan Jones. 1992. Syntactic context and the shape bias in children's and adults' lexical learning. *Journal of Memory and Language* 31, 801–825.

Landau, Barbara, Linda B. Smith & Susan Jones. 1998. Object perception and object naming in early development. *Trends in Cognitive Sciences* 2, 19–24.

Landau, Barbara & Deanna S. Stecker. 1990. Objects and places: Geometric and syntactic representation in early lexical learning. *Cognitive Development* 5, 287–312.

Landman, Fred. 1991. *Structures for Semantics*. Dordrecht: Kluwer.

Landman, Fred. 2000. *Events and Plurality: The Jerusalem Lectures.* Dordrecht: Kluwer.

Landman, Fred. 2004. *Indefinites and the Type of Sets*. Oxford: Blackwell.

Levelt, Willem J.M. 1989. *Speaking: From Intention to Articulation*. Cambridge, MA: The MIT Press.

Levinson, Stephen C. 2003. *Space in Language and Cognition: Explorations in Cognitive Diversity*. Cambridge: Cambridge University Press.

Levinson, Stephen C. & Sergio Meira. 2003. 'Natural concepts' in the spatial topological domain — adpositional meanings in cross-linguistic perspective: An exercise in semantic typology. *Language* 79, 485–516.

Lewis, David. 1986. *On the Plurality of Worlds*. Malden, MA: Blackwell.

Link, Godehard. 1983. The logical analysis of plurals and mass terms: A lattice theoretical approach. In Rainer Bäuerle, Christoph Schwarze & Arnim von Stechow (eds.), *Meaning, Use, and Interpretation of Language*, 302–323. Berlin: Mouton de Gruyter.

Link, Godehard. 1998. *Algebraic Semantics in Language and Philosophy*. Stanford, CA: CSLI.

Maier, Emar. (2006). Belief in context: Towards a unified semantics of de re and de se attitude reports. Nijmegen: Radboud University Ph.D. dissertation.

Margolis, Eric & Stephen Laurence. 1999. Concepts and cognitive science. In Eric

Margolis & Stephen Laurence (eds.), *Concepts: Core Readings*, 3–82. Cambridge, MA: MIT Press.

Marr, David. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, CA: Freeman and Co.

Marr, David & H. Keith Nishihara. 1978. Representation and recognition of the spatial organization of three dimensional structure. *Proceedings of the Royal Society of London B* 200, 269–294.

Mix, Kelly S., Linda B. Smith & Michael Gasser. 2010. *The Spatial Foundation of Language and Cognition*. Oxford: Oxford University Press.

Montague, Richard. 1973. The proper treatment of quantification in ordinary English. In Jaakko Hintikka, Julius Moravcsik & Patrick Suppes (eds.), *Approaches to Natural Language*, 221–242. Dordrecht: Reidel.

Munnich, Edward & Barbara Landau. 2003. The effect of spatial language on spatial representations: Setting some boundaries. In Dedre Gentner & Susan Goldin-Meadow (eds.), *Language in Mind: Advances in the Study of Language and Thought*, 113–155. Cambridge, MA: MIT Press.

Muskens, Reinhard. 1996. Combining Montague semantics and discourse representation. *Linguistics & Philosophy* 19, 143–186.

Nam, Senghou. 1995. The semantics of locative prepositional phrases in English. Los Angeles, CA: UCLA Ph.D. Dissertation.

O'Keefe, John. 1983. A review of the Hippocampus place cells. *Progress in Neurobiology* 13, 419–439.

O'Keefe, John. 1990. A computational theory of the hippocampal cognitive map. *Progress in Brain Research* 83, 301–312.

O'Keefe, John. 1991. The hippocampal cognitive map and navigational strategies. In Jacques Paillard (ed.), *Brain and Space*, 273–295. New York: Oxford University Press.

O'Keefe, John. 1996. The spatial prepositions in English, vector grammar, and the cognitive map theory. In Bloom *et al.* (eds.), 277–301.

O'Keefe, John. 2003. Vector grammar, places, and the functional role of the spatial prepositions in English. In van der Zee & Slack (eds.), 70–95.

O'Keefe, John & Neil Burgess. 1999. Theta activity, virtual navigation and the human hippocampus. *Trends in Cognitive Science* 3, 403–406.

O'Keefe, John & Lynn Nadel. 1978. *The Hippocampus as a Cognitive Map*. Oxford: Clarendon Press.

Parsons, Terence. 1990. *Events in the Semantics of English*. Cambridge, MA: MIT Press.

Phillips, Colin. 1996. *Order and Structure*. Cambridge, MA: MIT dissertation.

Piattelli-Palmarini Massimo & Juan Uriagereka. 2005. The evolution of the narrow language faculty: The skeptical view and a reasonable conjecture. *Lingue e Linguaggio* IV, 27–79.

Pinker, Stephen, & Ray Jackendoff. 2005. The faculty of language: What's special about it? *Cognition* 95, 201–236.

Poggio, Tomaso & Shimon Edelman. 1990. A network that learns to recognize three dimensional objects. *Nature* 343, 263–266.

Pylyshyn, Zenon W. 1984. *Computation and Cognition: Towards a Foundation for*

Cognitive Science. Cambridge, MA: MIT Press.

Pylyshyn, Zenon W. 1989. The role of location indexes in spatial perception: A sketch of the FINST spatial-index model. *Cognition* 32, 65–97.

Pylyshyn, Zenon W. 1994. Primitive mechanisms of spatial attention. *Cognition* 50, 363–384.

Pylyshyn, Zenon W. 2001. Visual indexes, preconceptual objects, and situated vision. *Cognition* 80, 127–158.

Pylyshyn, Zenon. 2003. *Seeing and Visualizing: It's Not What You Think*. Cambridge, MA: MIT Press.

Pylyshyn, Zenon W. 2004. Some puzzling findings in multiple object tracking (MOT): I. Tracking without keeping track of object identities. *Visual Cognition* 11, 801–822.

Pylyshyn, Zenon W. 2006. Some puzzling findings in multiple object tracking (MOT): II. Inhibition of moving nontargets. *Visual Cognition* 14, 175–198.

Pylyshyn, Zenon W. & Vidal Annan Jr. 2006. Dynamics of target selection in multiple object tracking (MOT). *Spatial vision* 19, 485–504.

Quine, William Orman van. 1960. *Word and Object*. Cambridge, MA: MIT Press.

Ramchand, Gillian. (2008). *Verb Meaning and the Lexicon: A First Phase Syntax*. Oxford: Oxford University Press.

Regier, Terry & Laura A. Carlson. 2001. Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology: General* 130, 273–298.

Regier, Terry & Mingyu Zheng. 2003. An attentional constraint on spatial meaning. *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*, 50.

Regier, Terry, Laura A. Carlson & Bryce Corrigan. 2005. Attention in Spatial language: Bridging geometry and function. In Laura A. Carlson & Emile van der Zee (eds.), *Functional Features in Language and Space: Insights from Perception Categorization and Development*, 191–204. Oxford: Oxford University Press.

Reiss, Charles. 2007. Modularity in the sound domain: Implications for the purview of universal grammar. In Gilliam Ramchand & Charles Reiss (eds.), *The Oxford Handbook of Linguistic Interfaces*, 53–80. Oxford: Oxford University Press.

Reynolds, Jeremy R., Jeffrey M. Zacks & Todd S. Braver. 2007. A computational model of event segmentation from perceptual prediction. *Cognitive Science* 31, 613–643.

Riesenhuber, Maximilian & Tomaso Poggio. 1999a. Hierarchical models of object recognition in the Cortex. *Nature Neuroscience* 2, 1019–1025.

Riesenhuber, Maximilian & Tomaso Poggio. 1999b. A note on object class representation and categorical perception. *AI Memos 1679*, CBCL 183.

Riesenhuber, Maximilian & Tomaso Poggio. 2000. Models of object recognition. *Nature Neuroscience* 3 (suppl.), 1199–1204.

Riesenhuber, Maximilian & Tomaso Poggio. 2002. Neural mechanisms of object recognition. *Current Opinion in Neurobiology* 12, 162–168.

Rosch, Eleanor. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General* 104, 192–233.

Samuels, Bridget. 2009. The structure of phonological theory. Cambridge, MA:

Harvard University dissertation.

van der Sandt, Rob A. 1988. *Context and Presupposition*. London: Croom Helm.

van der Sandt, Rob A. 1992. Presupposition projection as anaphora resolution. *Journal of Semantics* 9, 333–377.

Scholl, Brian J. 2001. Objects and attention: The state of the art. *Cognition* 80, 1–46.

Scholl, Brian J. 2007. Object persistence in philosophy and psychology. *Mind & Language* 22, 563–591.

Schwarzschild, Roger. 1996. *Pluralities*. Dordrecht: Kluwer.

Serre, Thomas, Lior Wolf & Tomaso Poggio. 2005. Object recognition with feature inspired by visual cortex. *Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 994–1000.

Shutts, Kristin & Elizabeth Spelke. 2004. Straddling the perception–conception boundary. *Developmental Science* 7, 507–511.

Smith, Linda B., Susan S. Jones & Barbara Landau. 1996. Naming in young children: A dumb attentional mechanism? *Cognition* 60, 143–171.

Smith, Linda B., Susan S. Jones, Barbara Landau, Lisa Gershkoff-Stowe & Larissa Samuelson. 2002. Object name learning provides on-the-job training for attention. *Psychological Science* 13, 13–19.

Soja, Nancy N., Susan S. Carey & Elizabeth Spelke. 1992. Perception, ontology, and word meaning. *Cognition* 45, 101–107.

Soschen, Alona. 2008. On the nature of syntax. *Biolinguistics* 2, 196–224.

Speer, Nicole K., Kehna M. Swallow & Jeffrey M. Zacks. 2003. Activation of human motion processing areas during event perception. *Cognitive, Affective, & Behavioral Neuroscience* 3, 335–345.

Spelke, Elizabeth S. & Susan Hespos. 2001. Continuity, competence, and the object concept. In Emmanuel Dupoux (ed.), *Language, Brain, and Cognitive Development: Essays in Honor of Jacques Mehler*, 325–340. Cambridge, MA: MIT Press.

Spelke, Elizabeth S., Gary Katz, Susan E. Purcell, Sheryl Ehrlich & Karen Breinlinger. 1994. Early knowledge of object motion: Continuity and inertia. *Cognition* 51, 131–176.

Spelke, Elizabeth S. & Gretchen A. van de Walle. 1993. Perceiving and reasoning about objects: Insights from infants. In Naomi Eilen, Rosaleen McCarthy & Bill Brewer (eds.), *Spatial Representation*, 110–120. Oxford: Basil Blackwell.

Stalnaker, Robert. 1973. Presuppositions. *Journal of Philosophical Logic* 2, 447–457.

Stalnaker, Robert. 1999. *Context and Content: Essays on Intentionality in Speech and Thought*. Oxford: Oxford University Press.

Stankiewicz, Brian & John E. Hummel. 1996. MetriCat: A representation for subordinate and basic-level classification. *Proceedings of the 18th Annual Meeting of the Cognitive Science Society*, 254–259.

Stockhof, Martin, Jeroen Groenendijk & Frank Veltman. 1996. In Shalom Lappin (ed.), *Coreference and Modality: Handbook of Contemporary Semantic Theory*, 179–213. Malden, MA: Blackwell.

Svenonius, Peter. 2006. The emergence of axial parts. *Nordlyd* 33, 49–77.

Talmy, Leonard. 1978. Figure and ground in complex sentences. In Joseph H. Greenberg, Charles E. Ferguson & Edith A. Moravcsik (eds.), *Universals of Human Language*, vol. 4, 627–649. Stanford, CA: Stanford University Press.

Talmy, Leonard. 2000. *Towards a Cognitive Linguistics*. Cambridge, MA: MIT Press.

Trofimova, Maria. 2009. Case assignment by prepositions in Russian aphasia. Groningen: University of Groningen Ph.D. dissertation.

Tulving, Endel. 1972. Episodic and semantic memory. In Endel Tulving & Wayne Donaldson (eds.), *Organization of Memory*, 382–402. New York: Academic Press.

Tulving, Endel. 1983. *Elements of Episodic Memory.* Oxford: Clarendon Press.

Tulving, Endel. 2000a. Memory: Overview. In Alan E. Kazdin (ed.), *Encyclopedia of Psychology*, vol. 5, 161–162. New York: American Psychological Association and Oxford University Press.

Tulving, Endel. 2000b. Introduction to memory. In Michael S. Gazzaniga (ed.), *The New Cognitive Neurosciences*, 2nd edn., 727–732. Cambridge, MA: MIT Press.

Tulving, Endel. 2002. Episodic memory: From mind to brain. *Annual Review of Psychology* 53, 1–25.

Tversky, Barbara, Jeffrey M. Zacks & Bridgette Martin Hard. 2008. The structure of experience. In Thomas F. Shipley & Jeffrey M. Zacks (eds.), *Understanding Events*, 436–464. Oxford: Oxford University Press.

Ullman, Shimon. 1979. *The Interpretation of Visual Motion*. Cambridge, MA: MIT Press.

Ullman, Shimon. 1996. *High Level Vision*. Cambridge, MA: MIT Press.

Vaina, Lucia M. (ed.). 1990. *From the Retina to the Neocortex: Selected Papers of David Marr*. Boston, MA: Birkhauser.

Walle, Gretchen van der & Elizabeth S. Spelke. 1996. Spatiotemporal integration and object perception in infancy. *Child Development* 67, 621–2640.

Winter, Yoad. 2008. Between logic and common sense — Or: Dikkertje Dap meets Dr. Halfbaked. Talk presented at the Nijmegen Institute for Cognition and Information, Nijmegen (20 May 2008).

Wunderlich, Dieter. 1991. How do prepositional phrases fit into compositional syntax and semantics? *Linguistics* 29, 591–621.

Zacks, Jeffrey M. 2004. Using movement and intentions to understand simple events. *Cognitive Science* 28, 979–1008.

Zacks, Jeffrey M., Nicole K. Speer, Kehna M. Swallow, Todd S. Braver & Jeremy R. Reynolds. 2007. Event perception: A mind/brain perspective. *Psychological Bulletin* 133, 273–293.

Zacks, Jeffrey M. & Kehna M. Swallow. 2007. Event Segmentation. *Current Directions in Psychological Science* 16, 80–84.

Zacks, Jeffrey M. & Barbara Tversky. 2001. Event structure in perception and conception. *Psychological Bulletin* 127, 3–21.

Zacks, Jeffrey M., Barbara Tversky & Gowri Iyer. 2001. Perceiving, remembering and communicating structure in events. *Journal of Experimental Psychology: General* 130, 29–58.

van der Zee, Emile. 2000. Why we can talk about bulging barrels and spinning spirals: Curvature representation in the lexical interface. In Emile van der Zee & Urpo Nikanne (eds.), *Cognitive Interfaces: Constraints on Linking Cognitive Information*, 143–184. Oxford: Oxford University Press.

van der Zee, Emile & Jon Slack (eds.) 2003. *Representing Direction in Language and Space*. Oxford: Oxford University Press.

Zwarts, Joost. 2005. Prepositional aspect and the algebra of paths. *Linguistics & Philosophy* 28, 739–779.

Zwarts, Joost. 2010. A hierarchy of locations: Evidence from the encoding of direction in adpositions and cases. *Linguistics* 48, 983–1009.

Zwarts, Joost & Yoad Winter. 2000. Vector space semantics: A model-theoretic analysis of locative prepositions. *Journal of Logic, Language and Information* 9, 169–211.

Francesco-Alessio Ursini
Macquarie University
Macquarie Centre for Cognitive Sciences (MACCS)
Building C5C, Talavera Road
2113 North Ryde (Sydney), NSW
Australia
francescoalessio.ursini@students.mq.edu.au