4 From Binding to Interpretation and Back

Language provides humans with finite means for generating an open-ended (or, theoretically, infinitely large) set of sentences (von Humboldt, 1836; Chomsky, 1965). In addition, each sentence may acquire a different meaning as a function of the contexts in which it is used. Factual and observation statements from the natural sciences (e.g., 'Ordinary neutrinos have small mass' and 'Jupiter has 67 moons') and what used to be called "analytic statements" (e.g., 'Bachelors are unmarried men') have essentially the same meaning regardless of the situations in which they are used and of the identities of the speaker and the addressee.[1] By contrast, deictic expressions, such as 'here' and 'now,' and indexicals, more generally, including many pronouns, adverbs, and adjectives, refer to someone or something different, depending on the context of usage (Kamp, 1971; Perry, 1977; Kaplan, 1989; Stalnaker, 2014). Productivity, or "the infinite use of finite means" (von Humboldt, 1836), is often considered as *the* hallmark of language. Yet, context sensitivity is equally characteristic, just as pervasive, and arguably no less formidable a challenge for theories of meaning in the brain.

The challenge comes from the fact that, typically, natural language presents a variable mixture of the crystallized semantics of factual and analytic sentences, on the one hand, and of the fluid referential behavior of indexicals, on the other. For example, the sentence

(1) John believes that Peter's latest book is extremely good.

involves relatively context-invariant meanings (e.g., 'believes' and 'book') that, when combined with other expressions, may result in partly context dependent meanings; for example, 'latest book' refers (implicitly) to the time at which the

[1] According to the official NASA website, "Jupiter, the largest planet in the solar system, has 67 moons and counting." The truth value of observation statements might well change, but that would not render their meaning any more context dependent. There would, however, be greater flexibility of interpretation if, for example, the definition of 'moon' became contested, and if alternative uses emerged (Caws, 1959; Achinstein, 1965; Nola, 1980; Papineau, 1996).

sentence is uttered.[2] The need to balance and blend context-dependent aspects of meaning and context-invariant ones arises at all levels of linguistic structure. Even seemingly "hard" logical expressions, such as propositional connectives and quantifiers, may shift in meaning, also depending on context. Here, I argue that the main computational goal of the I-system is to construct a representation (i.e., a "model") that makes discourse "true" and in which context-invariant and context-dependent aspects of meaning are balanced and blended. This exceeds by far the binding problem in semantics in a narrow sense. Given the sentences in figure 1.2, the I-system intervenes to determine who 'we' and 'they' refer to in a particular context of utterance, whether these sentences admit only literal or also figurative and pragmatic interpretations, what inferences may be drawn from them, and so forth. The binding problem is surrounded by a constellation of referential, elaborative, and inferential problems. A correct analysis of these problems is a critical test for a neuroscience of natural language use (Willems, 2015) as well as for theories of meaning in the brain.

What are the relations between semantic binding and interpretation, between the R- and I-systems? As argued in chapters 2 and 3, the amplitude of the main neural signature of relational processing, the N400, is inversely correlated with the degree of (semantic) fit between the eliciting word and the context in which it occurs. Crucially, the context could be *anything* that carries meaning, such as images, movies, gestures, single words, sentences, or discourse, irrespective of whether the relevant stimuli are syntactically organized (Kutas and Federmeier, 2011). *The I-system contributes to representations of the context into which the R-system binds word meanings* (van Berkum, 2004). That would explain N400 effects evoked by words whose meaning is a poor fit with the current *discourse* representation (van Berkum et al., 1999b, 2003b; Nieuwland and van Berkum, 2005b). If the N400 reflects binding based on stored knowledge *and* temporary discourse relations, it would seem unnecessary to posit an I-system, in addition to the R-system (Hagoort and van Berkum, 2007). We will discuss theoretical arguments and empirical evidence for a functional separation of relational and interpretive operations and for the view that processing dependencies between the R- and I-systems are not linear, but instead *circular*. Binding provides new input to interpretation, and interpreted relational structures constitute the input, with the incoming word, to the next stage of binding. To understand how labor is shared and divided between the R-system and I-system, at the computational and algorithmic levels, we should take a brief detour into formal semantics.

[2] Implicit reference is involved in numerous phenomena in semantics and pragmatics, such as verb tense and aspect, as Reichenbach (1947) has shown. However, there is little experimental research on the topic. I return to this issue later in this chapter and in chapters 5 and 6.

Formal semantics enables us to grasp the problem of meaning construction in the brain in its full extent and to articulate it with greater precision than would otherwise be possible (Baggio et al., 2012a,b, 2016). In chapters 1–3, we have argued that *composition* (bottom-up binding), as envisaged by semantic theory, has a correlate in the brain. However, (a) its scope must be regimented, because in some cases *contextual preactivation* (top-down binding) generates the same structures that composition would produce, given appropriate inputs, and (b) it should be viewed as a semantic operation that relies on (but does not reduce to) syntax, in accordance with autonomous semantics and in partial disagreement with compositionality. Certain restrictions also apply to *interpretation*, another fundamental concept of formal semantics. Meaning is assigned to expressions by means of *interpretation functions*. Logical forms are merely devices, utterly dispensable in many versions of Montague grammar (Montague, 1970a,b), for mapping natural language expressions onto reference structures, so as to satisfy compositionality. The interpretation function and the reference structure (plus, if required, further semantic elements) are jointly called a *model*. What would be a model for 'Dogs bark,' for example? Here, a model $\mathbf{M} = \langle D, i \rangle$ is a pair,[3] where D is the *domain* of interpretation (described in set-theoretic terms), and i is an *interpretation function* that assigns each individual constant (e.g., 'Fido') to an element of D (Fido), each predicate ('dog') to a subset of D (the set of all dogs), and so on. (Tarski, 1944). The meaning of 'Dogs bark' may be specified by its *truth conditions*, namely the class of *all* models in which dogs bark.

The model-theoretic analysis of meaning has had much success in linguistics and philosophy. However, it appears to be sharply disconnected from cognitive reality. First, it is unclear whether and how the brain can handle several models simultaneously, as seems to be required if meanings really are truth conditions. In chapter 1, it was argued that the human brain instantiates *one* representation at a time (e.g., of a perceptual object or a sentence). Moreover, representational stability (RS) applies to such "internal models." If that is true, the meaning that results from interpretation (a discourse representation) need not be a complete specification of truth conditions. Rather, it is a *single model*. Formal semantics in the Montagovian tradition, however, has little to say about such a "preferred model." 'Dogs bark' is true in a model in which there are only dogs (so long as they all bark); in a model with dogs, coyotes, and dingos, and they all bark; and numerous others. In fact, there are *infinitely many models*, as there are *infinitely*

[3] For simplicity, I assume that the meaning of 'Dogs bark' may be captured by standard first-order predicate logic; however, in chapter 1 we saw how it can be composed in a higher-order logic with λ-operators via function application. For an introduction to the formalism, see Gamut (1991) and Carpenter (1997).

many ways to construct a suitable domain (Gamut, 1991); hence the need for a *minimal model*—the model that posits the *smallest possible reference structure that makes the discourse "true"*. In the minimal model, there are no coyotes or dingos, only barking dogs. If one understands interpretation as the construction of a single minimal model, one needs a suitable notion of inference. In classical logic, deduction is *monotonic*. If a conclusion follows from the premises (e.g., 'Dogs bark'), then it also follows from an *extended* set of premises (e.g., 'Dogs bark, and coyotes and dingos do it too'). However, this monotonicity cannot be upheld in a formal system that computes minimal models (van Lambalgen and Hamm, 2004b; Stenning and van Lambalgen, 2005, 2008). In a minimal model of 'Dogs bark,' there are no entities other than dogs. In such a model, 'Coyotes and dingos do it too' is "false," in the technical sense that nothing about coyotes and dingos could be deduced from the model.[4] One has to allow *nonmonotonic* transitions between minimal models, so that sentences may change truth values as discourse proceeds.

4.1 Semantics as Computational Theory

Can conceptual insights and technical solutions from formal semantics be used to characterize computational-level theories of meaning in the brain? There are two arguments that would seem to undermine such a project from the start, but that do not stand up to closer scrutiny. The first is the historical observation that previous attempts to derive processing predictions from linguistic theories have failed, and there is no obvious reason why renewed efforts would be successful. One classic example is the Derivational Theory of Complexity (DTC). In DTC, processing costs were predicted to be a function of the number of steps required by the syntactic derivation of a particular sentence (Chomsky, 1965; Fodor and Garrett, 1967; Chomsky, 1968). In spite of initial support for DTC from results on passivization, subject-auxiliary inversion, and others (Mehler, 1963; Miller and McKean, 1964; Savin and Perchonok, 1965), subsequent research findings were mixed or contrary (Wason, 1965; Fodor and Garrett, 1966; Slobin, 1966; Fodor and Garrett, 1967; Forster and Olbrei, 1973).[5] Why did DTC fail, and what can one learn from its generally acknowledged failure?

[4] The domain of interpretation is a "closed world," which is completely described by the premises: what is not entailed by the premises is not the case (Clark, 1978; Reiter, 1978).

[5] It may be interesting to note that, in some of these studies, the evidence *against* the Derivational Theory of Complexity was simultaneously evidence *for* crosstalk between grammar and a (partly) autonomous semantics (Wason, 1965; Slobin, 1966).

Presumably, DTC failed because it assumed a *transparent mapping* between competence and performance: the parser implements syntactic operations (e.g., transformations) in a *serial fashion* (Berwick and Weinberg, 1983). Moreover, DTC did not bother to specify what should count, in a given theory of grammar, as a formal operation that *is* reflected in parsing. Culicover and Nowak (2002) distinguish between *core operations*, necessary to compute syntactic structure, and *housekeeping devices*, ad hoc solutions that are bound to be abandoned (or revised in newer versions of the theory) and are less likely to affect processing. Even with these caveats, the central insight of DTC can perhaps be maintained. Theories of meaning and grammar *can* constrain processing theories in various ways and to varying degrees (Poeppel and Embick, 2005; Phillips and Wagers, 2007; Sag and Wasow, 2011). As Marantz (2005) has observed, the notion that analyses of formal structure constrain processing models is of a piece with the standard methodology (possibly even the dominant methodology) in cognitive neuroscience. To avoid some of the pitfalls of early incarnations of DTC, it is useful to distinguish computational-level and algorithmic-level analyses, as an alternative to the distinction between "competence" and "performance" (Marr, 1982; Jackendoff, 2002; Baggio et al., 2012a). In his book *Vision*, Marr writes

> Chomsky's (1965) theory of transformational grammar is a true computational theory in the sense defined earlier. It is concerned solely with specifying what the syntactic decomposition of an English sentence should be, and not at all with how that decomposition should be achieved. Chomsky himself was very clear about this—it is roughly his distinction between competence and performance. (Marr, 1982, p. 28)

This holds for the I-system, too: a nonstandard (and stripped-down) version of semantic theory would specify what representations may be computed and the "logic of the computation;" the formal algorithmic theory would describe how the representations are set up; and finally, the neural theory would explain how the algorithms are implemented as patterns of activity in the brain.

A more radical argument against a constructivist approach of this kind is that formal semantics has *nothing to offer* for implementation, much like axiomatic theories of natural numbers say nothing about how humans might represent and process discrete quantities. The question posed by Partee (1980)—is semantics mathematics or psychology?—receives a definite answer here. There are a few similarities in how semantics and mathematics are pursued, such as their use of proofs (i.e., derivations) and model-theoretic techniques. Yet, these similarities belie rather different strategies and goals. In mathematics, a proof is a sequence of interlocking derivations that rely on a variety of methods (e.g., mathematical induction and contradiction) and of theoretical premises. The aim of a proof is to *expand the theory*, often to turn a conjecture into a theorem (Lakatos, 1976).

In semantics, by contrast, a derivation is the sequential application of a limited set of prespecified operations (e.g., λ-conversion or function application) to the same objects recursively, given specific constraints (e.g., compositionality) and premises. The goal of the derivation is to *validate the theory*, often to prove that the meaning of certain linguistic expressions may be derived (compositionally) and that the resulting interpretations conform to one's intuitions.[6] Only if the intended interpretations are derived and certain consequences follow (e.g., one gets the right synonymy and entailment patterns), can the theory be assumed to have the appropriate ontology and operations, until it gets tested on the next set of linguistic structures. This gradual effort to "get the logic right" is compatible with the aims of a computational theory in Marr's sense: to isolate the essential formal ingredients of a theory that can adequately describe semantic structures, given data provided by intuition, corpora, experiments, and other sources.

The historical and formalist arguments can thus be dismissed. Semantics can be construed as a cognitive modeling project, in which theoretical analyses can guide and constrain processing predictions at the algorithmic and neural levels (Baggio and van Lambalgen, 2007; Baggio et al., 2012a,b, 2015). But there is a difference between two requirements: (a) that (formal) semantics be pursued as a kind of computational-level theory and (b) that it be "realistic philosophically and psycholinguistically" (Hintikka, 1983, p. 20). Cognitive realism remains a hazy notion, because it relies on intuitions or philosophical arguments on what the human mind can (or cannot) represent. For example, high-order predicates, possible worlds, and types are occasionally dismissed as being less cognitively plausible or realistic than individuals and properties (Hintikka, 1983). It is then unclear why, historically, first-order and higher-order logics were developed at the same time, if the former are more "natural" than the latter (van Heijenoort, 2002), and why, in general, the human mind should be unable to handle entities that the logician's mind in particular can construct. There is a lack of solid data on the ontologies the brain can actually support. That is a fundamental problem that requires a concerted effort in logic, semantics, developmental psychology, and neuroscience. One could imagine cases where cognitively realistic theories are difficult to implement algorithmically or, vice versa, where theories that are less cognitively realistic are easier to implement and therefore to test. Realism may be a useful guide in generating novel hypotheses, but only the theories that can be implemented in specific processing and neural architectures are testable. Ultimately, the only way of knowing whether a supposedly cognitively realistic theory is empirically adequate is to test its processing consequences.

[6] This is not to deny that some (e.g., constructive) proofs in mathematics work that way. They are, however, less representative of mathematics as a whole than they are of formal semantics.

4.1.1 Computational Goals of the I-system

Formal semantics may therefore be used to characterize the computational goal of the I-system in discourse processing as the construction of a *minimal model* of the input. I will unpack and clarify this proposition as we proceed, because it conceals several subtleties and complications (e.g., the inputs are not sentences and models may not be iconic representations of reality). First, we must answer two preliminary questions: (a) is it useful to distinguish models from relational structures, as computed by the R-system via semantic binding? (b) How should models be understood—mentalistically or realistically? The short answers are: (a) yes; (b) mentalistically. Extended answers are given in what follows.

In a seminal work titled "Formal Semantics and the Psychology of Meaning," Johnson-Laird (1982) introduced a distinction between two kinds of (semantic) structure: *propositional representations* and *mental models*. The essence of his argument is that discourse comprehension involves two processing levels. At a more "superficial" level, senses (intensions) are activated and related—roughly what the R-system does here. The result is a propositional representation (PR): a structural analysis *of the given discourse*, and not of what discourse signifies. The PR is not a model, in the formal semantic sense, but rather the "description of a model," in Johnson-Laird's own words. PRs may suffice for understanding expressions whose meaning is context-invariant, or relies on stored knowledge, such as analytic sentences ('All canaries are birds') and some factual sentences ('Neutrinos have very small mass'). N400 data suggest that the R-system takes the lead in those cases (chapter 2). Johnson-Laird argues that one also needs a "deeper" level of representation, where mental models (MMs) are constructed, in order to explain how we understand context-dependent expressions. An MM is a cognitive representation *of what the discourse means*, of core aspects of the states of affairs it describes. Studies show that people often retain the "gist" of discourse, and only in specific circumstances do they encode the propositional form verbatim (e.g., if the discourse is indeterminate and does not allow for the construction of a definite model).[7] The relations between the present proposal and Johnson-Laird's analysis can then be summarized as follows: the R-system binds together semantic types and tokens into a "propositional representation" of the input, in the specific form of a *relational structure*; the I-system builds a "mental model" of the input, in the specific form of a *minimal discourse model*. An affirmative answer to (a) supports the distinction between the R-system and I-system in the brain. Experimental evidence is provided in chapter 5.

[7] See Bransford et al. (1972), Stenning (1978), Ehrlich et al. (1979), Johnson-Laird (1982), Mani and Johnson-Laird (1982), and Stenning (1986) for some empirical results. For recent discussions, see Stenning and van Lambalgen (2008) and Baggio et al. (2016).

Let us now tackle question (b): what is a suitable philosophical interpretation of models? The answer appears to hinge on how we choose to formalize mental models and on how we address the following conundrum: there is no easy way to fit truth conditions "into the head," and no guarantee that inconsistencies will be avoided when minimal models are embedded "into the world." As discussed earlier, there are infinitely many (classical) models for any given sentence (e.g., 'Dogs bark'), because there are infinitely many reference structures (domains) where its constituents may be satisfied: with just dogs; with dogs, coyotes, and dingos; and others. Such is the unwieldy truth-conditional meaning of a simple sentence (Gamut, 1991; Barba, 2007). One solution to fit (classical) structures "into the head" is to hold that what are in fact internalized are *finite procedures* for building models; that is, what are internally represented are not *all models*, but compact "intensional representations" in the form of algorithms computing *some models* (a finite number; e.g., one) on demand (Miller and Johnson-Laird, 1976; Johnson-Laird, 1982; Moschovakis, 1994; van Lambalgen and Hamm, 2004a).[8] *Realism is the only viable interpretation of classical models and truth conditional meaning*.[9] Conversely, trying to embed minimal models "into the world" (of classical models) leads to contradictions: a minimal model of 'Dogs bark' entails that nothing else is the case except that dogs bark; this information is filled in deductively by the logical system plus certain assumptions, and thus it may just not be true in situations in which coyotes and dingos bark. It follows that *mentalism is the only correct interpretation of minimal models*. Jackendoff (1990) has suggested that realism and mentalism in semantics are choices that depend on the purposes the theory is designed to fulfill. Here, I have reached a consonant conclusion. Realism and mentalism would follow from one's choice of models. To a large extent, this choice then dictates the aims that one's theory is able to serve, that is, philosophical analysis or cognitive modeling.[10]

---

[8] But how can classical intensions be "in the head" *and* determine extensions? This conjunction is precisely what Putnam's Twin Earth argument suggests is incoherent (Putnam, 1973, 1975). As a way out, one could assume that intensions only involve *partial functions*, so the structures that are computed are *partial models* (Partee, 1980; Muskens, 1995). Mentalism would then apply to *both* intensions *and* extensions. Johnson-Laird (1982) has argued it is impossible to know (whether we know) complete intensions: intensions are "philosophical idols" and as such empirically irrelevant.

[9] For further details on a realist account of Tarski's model-theoretic semantics (Tarski, 1944; Sher, 1999), see Niiniluoto (2004).

[10] One way to provide a uniform mentalistic framework for both classical and partial (or minimal) models would be, following Partee's (1980) suggestion, to view semantics in the tradition of Frege, Tarski, and Montague as the study of "super-competence," specifying what a mind could represent, if it were not limited by the finiteness of the brain; in Partee's own words, "a semantics for English as spoken by God." This proposal fails precisely where the finiteness of the human mind has to be factored in for the purposes of semantic analysis (e.g., with propositional attitudes; Partee 1980).

If the computational goal of the I-system is to compute a minimal (or partial) model of discourse, then how does the resulting theory explain what traditional formal semantics was conceived to explain, namely (a) how language connects to the world and (b) how people may understand each other. If minimal models cannot be embedded in the world, lest they collapse under the weight of logical contradiction, how is the reference structure (an internal representation) related to reality? The structure itself cannot be interpreted model-theoretically: only a language can. So, the relation between the structure and reality is *not* semantic. If intensions involve partial functions and models are either minimal or partial, then both can be internal representations: senses are algorithms, and references or reference structures are sets of values, returned by the algorithms.[11] Here, I assume that these internal representations are grounded in neural processes that get functionally connected across the R- and I-systems during the construction of a model of discourse. Within the I-system, the cortical processes that ground a reference structure may only be connected to reality *causally*. However, those connections are diverse. They are mediated by memory (e.g., of past entities or events), reasoning (abstractions), or perception (entities in one's environment) (Miller and Johnson-Laird, 1976; Jackendoff, 2002). *There is no single general answer to the question of how language is related to the world.* In a mentalistic framework, in which senses and references are partial structures, the relation is *causal*, and not semantic. It would be completely accounted for by physiology, if we did bother to reconstruct the full history of sensory stimulation and neural encoding for individuals in similar conditions (Quine, 1960).

Are minimal models an adequate basis for a computational-level analysis of discourse comprehension? Would partial models overlap to a degree sufficient for people to understand each other? Partee (1980) formulates the problem as follows:

> Finiteness restricts us to constructing partial models, and in place of complete intensions of words we construct imperfect algorithms which yield partial functions on these partial models. Different individuals will have different partial models and different algorithms, since our brains and our real-world experience are not identical. Communication will be possible as long as there is sufficient similarity in our partial models and our imperfect semantics. (p. 3)

[11] Moschovakis (1994) is careful to distinguish between *programs* and *algorithms*: "programs" are pieces of text, or sequences of instructions; "algorithms" are mathematical objects that interpret the programs. Not all procedural accounts of meaning are algorithmic in this sense. See, e.g., Pietroski (2008, 2011) for a nonalgorithmic theory of lexical meanings as instructions to retrieve or construct semantic representations. In general, relational and algorithmic semantics may be reconciled if one assumes that algebraic operations on lexical semantic vectors or the relata themselves (the vectors) may be interpreted algorithmically. Here, I endorse the first option, and I remain agnostic in regard to the second. I discuss some examples of algorithmic analyses of meaning in sections 4.2 and 4.3.

If Partee is right, formal semantics based on partial models would explain *both* successful *and* failed communication. But what, exactly, explains the fact that communication is *normally* successful? A partial answer is minimal models.[12] The speaker and the hearer may build different (nonminimal) models of 'Dogs bark,' but their internal representations cannot be models of 'Dogs bark' unless they overlap (at least) on the contents of the minimal model in which dogs bark, that is, on the smallest possible reference structure that would verify the phrase. Clearly, natural language use involves expressions whose intended meaning or message cannot be captured by a minimal model, and it requires nonminimal or partial models instead (examples to follow). Hence, in briefest outline, the goal of the I-system is threefold: (a) to construct a *minimal model of discourse*, what I will call *referential processing*; (b) to construct an *enriched discourse model*, that could implicate nonminimal or partial structures (*elaborative processing*); (c) to explore models, fleshing out implicit content, at the level of the reference structure, and to recompute minimal models, if needed (*inferential processing*). Many classic phenomena in semantics and pragmatics belong to either of these classes. Pronominal and temporal anaphora are cases of referential processing. Figurative meaning and forms of coercion ("logical metonymy") (Pustejovsky, 1995) involve elaborative processing. Finally, inferential processing subserves deductive, abductive, relational, and other forms of discourse-based reasoning, including some forms of pragmatic inference (e.g., presupposition denials and implicature cancellations). Experimental results are discussed in chapter 5.

What kind of semantic theory could provide the appropriate analytic tools to characterize the computational goals of the I-system? Discourse representation theory (DRT) (Kamp, 1981; Kamp and Reyle, 1993) has several properties that render it suitable for cognitive modeling. It is not possible to give a satisfactory account of the motivation of DRT and of its formal apparatus here. However, in brief, the theory was originally conceived to handle certain cases of pronominal and temporal anaphora that had proved recalcitrant to formalization in standard predicate logic. These cases are rather instructive in the present context, as they reveal fundamental facts about composition and interpretation that ought to be addressed by a computational and algorithmic theory of the I-system. Consider the minidiscourse:

(2) A dog barks. It is black.

[12] A satisfactory answer must await chapters 7–9. Understanding coordination and communication requires a computational theory in which speakers *and* hearers are modeled. Tracking coordination in this domain is the task of the E-system. If we consider semantics as a theory of human discourse comprehension, the "intended meaning," to which the product of interpretation should correspond, must be given by the intuition of the theoretician, as formal semantics lacks a model of the speaker.

The logical form of the first clause in (2) can be derived in a system of predicate logic with λ-operators, along the lines of 'Dogs bark,' as was shown in chapter 1. The result would be

(3) $\exists x[dog(x) \wedge bark(x)]$

Merely adding a conjunct for the second sentence, however, leads to a formula where one occurrence of x is not bound by the existential quantifier:

(4) $\exists x[dog(x) \wedge bark(x)] \wedge black(x)$

This fails to capture the anaphoric meaning of 'it.' To get the right translation

(5) $\exists x[dog(x) \wedge bark(x) \wedge black(x)]$

one has to modify (3) by undoing its brackets, so that $black(x)$ may be inserted inside the scope of the existential quantifier. If this maneuver were allowed, it would violate compositionality. Modifying the "bracket structure" of a formula is a syntactic operation without a definite counterpart in semantics. In standard predicate logic, (a) *composition is not a process* that may generate intermediate logical forms, so discourse must be translated as a whole, and (b) *interpretation is not a process* either and must wait for composition to produce the full logical form of discourse before it may be applied. Clearly, these are not very pleasant consequences from a psychological perspective.

The key innovation behind dynamic theories of meaning is to make discourse referents "accessible" *across* sentential boundaries. In dynamic predicate logic (DPL) (Groenendijk and Stokhof, 1991), values stick to variables until they are reset. This requires that quantifiers and quantifier scope be handled differently in DPL than in standard predicate logic. In DRT, referents would be accessible within a discourse representation structure (DRS). Discourse referents in DRT are unlike bound variables in predicate logics, as their behavior is not restricted by the syntax of quantifiers. The DRS for the first sentence of (2) is

(6) $[\,\mathbf{x} : dog(x), bark(x)]$

which specifies the conditions, introduced incrementally in discourse, applying to the discourse referent ($\mathbf{x}$), namely that it is a dog and that it barks. The clause 'It is black' introduces a new DRS, and (temporarily) a second referent ($\mathbf{y}$):

(7) $[\,\mathbf{y} : black(y)]$

Discourse integration here amounts to merging DRSs (6) and (7) by tentatively unifying the relevant variables, assuming a single referent in the final DRS:

(8a) $[\,\mathbf{x},\mathbf{y} : dog(x), bark(x), black(y)] =$
(8b) $[\,\mathbf{x},\mathbf{y} : x = y, dog(x), bark(x), black(y)] =$
(8c) $[\,\mathbf{x} : dog(x), bark(x), black(x)]$

The DRSs may be understood as a level of representation intermediate between syntax and the discourse model, which may not exactly reflect the logical forms of the sentences constituting the discourse. In this sense, DRT is often referred to as a "representational" and "noncompositional" theory of meaning, but DRT is essentially equivalent to DPL, which *is* compositional and does *not* involve a representational middleman between syntax (or logical form) and the discourse models (Groenendijk and Stokhof, 1991). What distinguishes DRT from other brands of (dynamic) semantics is neither its putative representationalism nor its repudiation of compositionality,[13] but rather (a) its usage of partial models, (b) its non-truth-conditional notion of meaning, and (c) its established equivalence to descriptively powerful versions of conceptual semantics (Jackendoff, 1983).

In DRT, models involve *partial structures*, which may, in general, and unlike minimal models, be embedded "into the world." DRT's partial models imply a different philosophical account of meaning than truth conditions. Consider the following sentences (Geurts and Beaver, 2011):

(9a) A dog barks.

(9b) It's not the case that a dog doesn't bark.

(9c) A dog barks and either it's snowing or it isn't.

In Montague grammar, these sentences share the *same truth conditions* and the *same meaning*: the set of all models in which a dog barks. In DRT, they all have *different meanings*, as the DRSs that are built up and interpreted differ in each case. In the DRS for (9a), the referent 'a dog' may be picked up anaphorically, as in (2), but negation in (9b) makes it less accessible. In addition, the DRS for (9c) contains sub-DRSs for each disjunct 'It is snowing' and 'It isn't snowing.' DRSs are richer structures than logical forms in Montague grammar and DPL. Zwarts and Verkuyl (1994) have shown that DRSs are essentially equivalent to Jackendoff's conceptual structures (Jackendoff, 1976, 1983, 1990, 1996): they are intertranslatable, and both can be interpreted model-theoretically via partial functions. Computation in the I-system is best captured by a formal framework that embodies the basics of DRT and related systems (Hamm et al., 2006), with one qualification: the representations computed by the I-system are, by default and at least initially, minimal models (van Lambalgen and Hamm, 2004b).

[13] Muskens (1996) examines mergers of DRT and Montague grammar in which compositionality is maintained, for example by adding λ-operators to DRT. There are various reasons, independent of discussions in post-Montagovian semantics, not to regard compositionality as a (strong) constraint for the R- and I-systems (recall chapter 1). Moreover, in principle, it is possible to eliminate DRSs altogether, transferring the representational apparatus of DRT to the models. One would then work directly with denotations, doing away with discourse referents entirely (Giorgolo and Unger, 2009; Geurts and Beaver, 2011).

4.1.2 Computational Logic of the I-system

The motivations for DRT suggest there are reasons, internal to semantic theory, for assuming that the goal of interpretation is to construct a model of *discourse*, transcending sentence boundaries. Moreover, the primacy of discourse over the sentence, as a domain of interpretation, has been emphasized in the psychology and neuroscience of language. Empirical evidence is discussed in chapter 5.[14] Besides identifying the main *goal* of I-system computation, a top-level analysis requires us to specify the *logic of the computation*, namely the characterization of the formal systems that best capture the relations between the inputs and the outputs of processing within the I-system. This requires addressing two issues: (a) Is the mapping *compositional*, that is, a (strict) function of the meanings of the constituents of discourse and of its syntactic structure? (b) Is it *monotonic*, that is, preserving the models computed at each successive stage?

Compositionality Three definitions of compositionality were quoted in chapter 1 (Partee, 1975; Partee et al., 1990; Partee, 1995). The third definition is:

> COMPOSITIONALITY 3 *The meaning of a whole is a function of the meanings of the parts and of the way they are syntactically combined.* (Partee, 1995)

According to the definition, compositionality is a soft constraint indeed. It falls to theories of meaning and grammar to determine precisely the semantics of the constituent parts, the syntactic mode of combination, and critically the relevant "function." Whether compositionality in fact holds or not should not depend on these choices. Yet, compositionality may *always* be shown to hold, if sufficient complexity (of the right kind) is built into lexical semantics or syntax (Janssen, 1986; Zadrozny, 1994; Kazmi and Pelletier, 1998; Westerståhl, 1998; Krifka, 1999). A related problem is that theories of syntax and of lexical semantics are usually provided *in the service of* compositionality, instead of being motivated independently (Groenendijk and Stokhof, 1991). Unless we *require* theories to be independently justified (for example, on purely formal grounds), processing constraints as revealed by research on language comprehension may also count as evidence for (or against) compositionality (Baggio et al., 2012b). Adjusting theories of syntax or lexical semantics to suit or to rescue compositionality has definite processing consequences, as the effects of theoretical choices percolate downward to the algorithmic and neural levels. It is essential to determine what these cognitive and neural constraints are, and in which cases compositionality collides with them. Some examples are discussed in what follows.

[14] For reviews and syntheses of findings in cognitive neuroscience, in particular from EEG studies, see van Berkum (2004, 2008, 2012). For the state of the art of experimental research on discourse comprehension prior to the diffusion of brain imaging, see Kintsch (1994), Carpenter et al. (1995), and Graesser et al. (1997).

A familiar argument for compositionality, the "productivity argument," starts from a perceived tension between the infinity of language and the finiteness of the brain. Substituting any large finite number for infinity here does not change the essence of the productivity argument, which is that not every sentence that may be produced and understood can be stored.[15] Compositionality is taken to be a solution to this problem. If meaning is generated compositionally, then the meaning of any complex expression may be *finitely* generated (Katz and Fodor, 1963). This has the (seemingly desirable) effect of lifting the burden on storage and placing it on computation, more precisely on the syntax. Compositionality apparently reconciles infinity and finiteness, language and the brain. However, upon closer inspection, placing greater weight on syntactic computation would not render one's processing theory any more brain-friendly, but rather quite the opposite. In part I, we saw how reliant the human brain is on stored (semantic) relations between representations of words and how this fact enables top-down binding through contextual preactivation. The N400 effect is striking evidence of this. One can then demonstrate that the processing consequences of shifting the costs of composition to syntax are not always aligned with empirical data.

One instructive example is provided by complement coercion. According to standard analyses (Pustejovsky, 1993; Pylkkänen and McElree, 2006), 'began' in (10) requires an *event* as a semantic argument, as in 'began writing the book' or in 'began the fight,' but not necessarily a VP as a syntactic complement:

(10) The journalist began the article before the coffee break.

The complement 'the article' denotes an *entity*. There are at least two strategies to circumvent the semantic mismatch between the verb's requirements and the complement's meaning. One is to assume a separate *semantic* operation, which takes the form of either *type shifting* (the semantic type of 'the article' is raised from the type of individuals to the type of predicates or events, so that ordinary composition and interpretation may apply) (Pustejovsky, 1993; Pylkkänen and McElree, 2006) or *unification* (an activity variable in the lexical representation of 'to begin' is assigned a contextual value, such as *writing* or *typing*) (Baggio et al., 2010). However, neither of these operations is reflected in the syntax of (10), therefore compositionality is violated.[16]

[15] For discussions of the distinction and balance between storage and computation in the language system, see the contributions collected by Nooteboom et al. (2002) and the recent computational and algorithmic analysis by O'Donnell (2015).

[16] Pustejovsky (1993) suggests that, if the trigger of type shifting is lexically specified on the verb, then compositionality may be preserved. This solution seems incompatible with a strong version of compositionality (syntax-semantics homomorphism, or rule-to-rule principle), because a semantic operation is triggered that has no counterpart in the syntax (Partee, 1975; Janssen, 1986, 1997).

Another strategy, which would preserve compositionality, consists precisely in assuming that the syntax of (10) includes a node for a *silent* lexical item (α). That would effectively "insert" a VP into the syntactic structure, corresponding semantically to the process or event satisfying the requirements of 'began:'

(11) [$_{NP}$ The journalist [$_{VP}$ began [$_{VP}$ α [$_{NP}$ the article]]]]

The linguistic evidence is mixed. There are arguments for (e.g., unaccusatives) and against (e.g., passivization) VP insertion (Pylkkänen and McElree, 2006). This would suffice to rule out, on strictly linguistic grounds, VP insertion as the correct analysis of complement coercion. Alternatively, if one assumes that the case cannot be decided based on linguistic intuition, one can turn to processing data for a resolution. As was observed in chapter 3, the computations triggered by 'article' in (10) are reflected by a larger N400 compared to the noncoercing condition, in which the verb is instead 'wrote' (Baggio et al., 2010; Kuperberg et al., 2010). In our ERP study, the N400 lasted longer than its typical duration. It is difficult to reconcile this N400 with VP insertion, which predicts instead a syntax-related ERP effect, such as a P600. This clearly shows that a theoretical move designed to rescue compositionality may lead to wrong predictions at the algorithmic and neural levels. Besides the theoretical reasons discussed earlier, there are therefore empirical grounds for assuming that compositionality might *not* hold in general and cannot be included in the computational logic of either the R-system or the I-system.

Monotonicity In classical logics, and in some other formalisms, often referred to as "nonclassical logics," inference is monotonic:

> MONOTONICITY *If a set of premises or discourse D licenses the inference that p (a proposition), then p also follows from any discourse D′ containing D.*

If 'Fido barks' follows from 'Dogs bark,' then it also follows from 'Dogs bark and cats meow.'[17] In a classical logic, 'Fido barks' follows from 'Dogs bark' if and only if 'Fido barks' is true in *all models* of 'Dogs bark.' That would make it true in the subset of models of 'Dogs bark' in which dogs bark and cats meow. Monotonicity flows naturally from a system in which interpretation establishes a *complete mapping* between language and reality. Therefore, if a single model can be produced in which dogs bark, but Fido does not, the system would block the inference 'Fido barks' from 'Dogs bark.' Problems with monotonicity arise when the system is not aware in advance of the existence of counterexamples. Suppose we are only allowed to model *fragments of the universe*. If we cannot

[17] One need not assume that Fido is a dog: it could be a dingo. Entailment patterns are not entirely constrained by categorial structures. Hence, the scope of inference in the I-system is not restricted to the relational structures stored within or computed by the R-system.

survey all models of discourse, then we cannot use the classical notion of valid inference: that p follows from discourse D if and only if p is true in *all models* of D. Here, an alternative definition of validity is required. In a nonmonotonic logic, p can follow from D even if p is true in a *single model* of D (the minimal model). However, this inference may be retracted, if required by the extended discourse D' (van Lambalgen and Hamm, 2004b).

To appreciate this point, consider again progressive clauses from chapter 1:

(12) The woman carrying a shopping bag was crossing the street.

It has been shown that the meaning of the VP 'to cross the street'—whether it is stored as a construction or computed online—may be captured by a minitheory that specifies the causal and temporal relations between different aspects of the eventuality, such as that crossing is completed only when the distance traversed by the woman equals the street's width (Hamm and van Lambalgen, 2003; van Lambalgen and Hamm, 2004b). The minitheory is effectively an *algorithm* for computing models. However, it requires an appropriate logic to deliver definite results (Hamm and van Lambalgen, 2003; van Lambalgen and Hamm, 2004b). There is an open class of events that may prevent the woman from crossing the street. To determine whether she will get to the other side, one has to be able to say something definite about adverse events. Unless the discourse implies that they do not happen, one cannot just rule them out. Classical logics contemplate models that include such obstacles, and others that do not. In terms of classical validity, nothing can be said as to whether she did actually cross the street.[18]

However, if we assume that sentence (12) describes a "closed world," and we include a principle of inertia in the system, then we may rule out occurrences of events that prevent the woman from reaching the other side. Now, the inference that she has crossed the street goes through (Hamm and van Lambalgen, 2003; van Lambalgen and Hamm, 2004b). Experimental data show that people draw inferences to the goal state when presented with progressive constructions such as (12) (Baggio and van Lambalgen, 2007; Baggio et al., 2008). The proportion of goal-state inferences drops—from 75% to 37% (Baggio and van Lambalgen, 2007) and from 76% to 55% (Baggio et al., 2008)—if discourse continues with a description of a disabling event, as in

[18] This inferential impasse is related to a major problem, known in the AI community as the "frame problem" (McCarthy and Hayes, 1969; McDermott, 1987); namely, how can one restrict reasoning to what is positively known, so that valid inferences about the effects of events may be drawn? This kind of problem is best seen as an illustration of the limitations of particular logical formalisms, as applied to human reasoning, rather than as an information processing problem faced by reasoners. Much as in the case of discourse anaphora, discussed here, the frame problem highlights important shortcomings of classical predicate logic as a tool for cognitive modeling.

(13) The woman carrying a shopping bag was crossing the street when she was hit by a motorcycle.

When the progressive clause is interpreted, "closed-world reasoning" allows us to exclude that the woman was hit by a vehicle and any other eventualities that cannot be derived from discourse or background knowledge. So, if one were to query the system, it would answer that such underivable statements are "false." This initial minimal model, where the woman has safely reached the other side of the street, must be recomputed when the subordinate clause comes in. Then, it is "true" (part of the *revised* minimal model) that she was hit by a motorcycle: that is sufficient to withdraw the inference that followed from the initial model. The transition between the initial and the revised models is nonmonotonic: p ('The woman has reached the other side of the street') follows from discourse D (12), but not from the extended discourse D' (13). Some inferences remain valid throughout the progression of a discourse, but interpretation in general is nonmonotonic (Baggio and van Lambalgen, 2007; Baggio et al., 2008).[19]

4.2 Principles of Discourse Processing

In the previous section, semantic computation in the R- and I-systems was seen as the noncompositional and nonmonotonic construction of a (minimal) model of discourse. Discourse comprehension involves the internal representation of the smallest possible reference structure that renders the discourse true (Zwaan and Radvansky, 1998; van Lambalgen and Hamm, 2004b). Setting aside some of the technical details discussed earlier, the computational analysis developed so far is adequately summarized by a statement due to George A. Miller:

> MILLER'S LAW *In order to understand what another person is saying, you must assume it is true and try to imagine what it could be true of.* (Hall, 1980)[20]

In what follows, we will move one level down in Marr's hierarchy, to provide a general characterization of the algorithms computing models in the I-system. I will not discuss specific algorithms in much detail, for two reasons. First, if the idea of senses as algorithms (Partee, 1980; Johnson-Laird, 1982; Moschovakis, 1994; van Lambalgen and Hamm, 2004a) is correct, each type of expression or construction in a given language must be associated with a different algorithm. What matters for our purposes are *invariant properties of classes of algorithms*

[19] For a discussion of forms of monotonicity in discourse processing, see Sturt and Crocker (1996, 1997, 1998). For experimental results on semantic reinterpretation, see Sturt (2007).

[20] This statement echoes Johnson-Laird's theory of mental models, and indeed these concepts have a shared ancestor in chapter 3 of *Language and Perception* (Miller and Johnson-Laird, 1976). For a history of mental models (i.e., a partial history of I-system theorizing), see Johnson-Laird (2004).

(corresponding to classes of expressions and constructions), which I refer to as *processing principles*. Second, the same meaning can be encoded in alternative grammatical forms in different languages and by different speakers (chapter 1), and processing algorithms may differ accordingly. The principles posited here apply at least to "canonical forms," or the simplest possible encoding of a given meaning into a syntactic form (i.e., the "syntactic analog" of a minimal model) (Gorrell, 1995; Culicover and Jackendoff, 2005, 2006; Culicover, 2013b,a).

First, consider the processing principles that would be suggested by the view of R- and I-system computation considered in the preceding section. That view is based on classical model theory and truth conditions (as a general account of interpretation) and on compositionality and monotonicity (as constraints on the mapping between syntax and semantics and on inference, respectively). In that view, composition (bottom-up binding) precedes and feeds into interpretation. There is a sequence of two steps with a fixed temporal relation. Neither process is incremental. Discourse is translated as a whole. Interpretation can be applied only after composition has resulted in a complete logical analysis of discourse. In addition, the whole process is modular and deterministic. If the discourse is well formed, it returns one or more logical forms (e.g., in the case of quantifier scope ambiguity): by virtue of compositionality, each logical form corresponds to a class of models. Processing is bottom-up, driven by syntax and the lexicon, and is never predictive.

In light of recent advances in psycholinguistics and neuroscience, the picture above seems utterly inadequate as a theory of human discourse comprehension. However, those are the processing consequences of classical formal semantics, if taken as a computational-level theory. That was effectively the starting point of computational semantics, which, however, moved away from the view above (Blackburn and Bos, 2005) by gradually relaxing or revising some of its central assumptions (e.g., allowing for incremental composition and interpretation, for nonmonotonic inference, partial or underspecified representations, and the use of background knowledge in computing models) and by integrating logical and statistical approaches to representation and inference.[21] In this section, I draw from the computational semantics tradition and from basic knowledge of brain function to characterize the processing logic of the I-system, under the guiding assumption of continuous and circular interactions between binding (including preactivation and composition) and interpretation.

[21] For overviews of progress in computational semantics, see the series of volumes on *Computing Meaning* (Bunt and Muskens, 1999; Bunt et al., 2001; Bunt and Muskens, 2007; Bunt et al., 2014). See the reviews by Baroni (2013), Liang and Potts (2015), and Stone (2016) for newer approaches, including vector-space semantics and distributional semantics.

4.2.1 Synchronous and Asynchronous Incrementality

The "dynamic turn" in semantics has led to the development of formal systems, such as DRT and DPL, where clauses and sentences are effectively interpreted incrementally (van Eijck, 2001). Incremental processing actually occurs below clausal boundaries (e.g., down to sublexical units) (Marslen-Wilson, 1973; Just and Carpenter, 1980; Altmann and Steedman, 1988). In the R-system, this kind of radical incremental processing, as evidenced by the N400, seems especially prominent, and arguably more so than in the I-system. The moderate version of incrementality, implemented in dynamic semantics, is not the result of intrinsic limitations of the formal apparatus. Rather, in dynamic semantics there is little need to posit word-by-word incrementality to capture the linguistic phenomena of interest (e.g., pronominal and temporal anaphora). This view of incremental processing may be enriched via the design of cognitively and neurally plausible clause- and sentence-level processing algorithms.

The key distinction is not between serial and parallel processing, as it is clear that there is massive parallelism in the neocortex, but between *synchronous* and *asynchronous* computation (Milner, 1983; Shieber and Johnson, 1993). Take a circuit of several interconnected elements. Each computes a particular function on the basis of the input it obtains. In a synchronous system, each element must wait for inputs to be available to other elements before it may begin to compute its own output. Computation across circuit elements, therefore, is at least partly coordinated (synchronized). In an asynchronous system, instead, each element begins to compute its output independently as soon as possible, that is, as soon as its inputs are sufficient to determine an output. It is likely that the brain relies on each architecture for different purposes. The principle of temporal evolution (TE; chapter 1), the hypothesis that binding and interpretation occur at multiple timescales, predicts that both "processing modes" are required. Asynchronous processing explains fast computing of the kind implemented in associative and attractor-based neural networks (Hopfield, 1982; Smolensky, 1986; Smolensky and Legendre, 2006) subserving R-system processes (chapter 3). Synchronous computation accounts for the slower, "punctuated" dynamics that characterize the I-system. Examples and evidence are presented here and in chapters 5–6.

The role of synchronous computation can be better appreciated by examining a concrete processing problem. Consider the following pair of sentences:

(14) Before/After the scientist submitted the paper, the journal changed its policy.

Münte et al. (1998) have shown that 'before' sentences give rise to long-lasting frontal negative ERPs spanning both clauses relative to 'after' sentences. What could be the sources of sustained computation here? Semanticists have tried to

formalize the observation that 'before' sentences do not entail the "truth" of the subordinate clause, whereas 'after' sentences do (Anscombe, 1964; Cresswell, 1977; Åqvist et al., 1978; Landman, 1991; Beaver and Condoravdi, 2003; del Prete, 2008). With 'before,' the scientist may have submitted the paper, or may not have done so, perhaps as a consequence of the journal's policy change. This nonveridical interpretation is not available with 'after' clauses.[22] Only 'before' clauses may be interpreted veridically or nonveridically, largely depending on the main clause, in particular on whether the event it describes has the potential to impede the occurrence of the event described in the subordinate clause. This knowledge is assumed to be made available by memory or the context.

There are two alternative processing strategies (and classes of algorithms) for interpreting 'before' and 'after' sentences. One is based on the incremental and nonmonotonic construction of a minimal model of the subordinate clause. The initial veridical interpretation may be revised while processing the main clause. But this account is at odds with the ERP data (Münte et al., 1998), which show an effect already at the first clause. Why would that be the case, if both 'before' and 'after' clauses are (initially) interpreted veridically? The second strategy is based on synchronous computation. If e_1 is 'The journal's policy change,' e_2 is 'The scientist's submission of the paper,' and if the relevant temporal ordering is $e_1 < e_2$, then the interpretation of 'Before e_2, e_1' is given by this algorithm:

(i) put event e_2 in a temporary store;

(ii) update the discourse model with e_1 situated in the past of *now* and mesh with already present events;

(iii) compute states consequent on e_1;

(iv) check whether any of these consequent states conflicts with conditions for the occurrence of e_2;

(v) if so, empty the store, but do not update the discourse model with e_2;

(vi) if not, empty the store and update the discourse model with e_2, situated between e_1 and *now* ($e_1 < e_2 < now$).

The I-system evaluates e_1 prior to e_2 (synchronous computation). By contrast, with 'after,' e_1 and e_2 may be immediately integrated into the discourse model:

(i) update the model with event e_1 and time t in the past and mesh e_1 with already present events;

(ii) update the resulting model with event e_2 and time s with $t < s$.

[22] One should distinguish between nonveridical (nonentailing) and antiveridical readings (Zwarts, 1995; Giannakidou, 1998). Only in the latter case would the negation of the subordinate clause be entailed by a 'before' sentence, plus a context and background knowledge. Veridicality and related notions may also be applied to expressions (e.g., 'before' is nonveridical, negation is antiveridical).

In 'Before e_2, e_1,' the event e_2 is held in (working) memory until relatively late before it can be integrated into the model. The word 'before' would act as a cue for the I-system to start storing e_2 (or perhaps a propositional representation of the first clause) in a short-term buffer. Could this explain the slow brain waves triggered by 'before,' increasing for both clauses, and the correlations between ERP amplitudes and participants' working memory (Münte et al., 1998)?

The preceding discussion suggests that synchronous computation requires a processing model with more assumptions. If there is no short-term storage, the I-system is forced to interpret its R-system inputs immediately and possibly to recompute the resulting model if necessary. Synchronous computation requires at least two additional components in its underlying architecture: (a) a working memory and (b) a system of communication across network units, so that each element receives a "go" signal when others are ready to start computing or have completed their task, or a functionally equivalent process.[23] This architecture is required by the kind of incrementality posited by DRT and DPL, in which the relevant processing chunks are phrases and clauses, and there exist interpretive dependencies between them and multiple options for ordering them at the point of encoding. Conditionals are a clear example ('If p then q' vs 'q if p'), and so are 'before' sentences. The principle of *synchronous incrementality* is required to account for computation in the I-system, while *asynchronous incrementality* applies to the R-system's functions and comprises both *effective incrementality* (e.g., sublexical and word-by-word) and *immediacy* (Levelt, 1978; Tanenhaus et al., 1995; Hagoort, 2003, 2006b; Hagoort and van Berkum, 2007; Kutas and Federmeier, 2011), or rapid asynchronous processing, together with the ability of the R-system to integrate information from different streams and modalities. As was shown in chapter 2, empirical results support this view of the R-system. Evidence of synchronous interpretive processing is provided in chapter 5.

If the formula R-system : asynchronous processing = I-system : synchronous processing is correct, then three predictions follow for the I-system in the brain: (a) interpretation requires partial input (not necessarily full sentences) from the R-system, so its ERP correlates should follow the N400; (b) these ERPs should track computation at longer timescales (phrases or clauses), so they should last longer and may resemble the slow waves associated with working memory; (c) they should appear at critical "interpretation points" in sentences or discourses (*punctuated processing*), at which reference is computed, models are extended and enriched, and inferences are drawn or withdrawn. These predictions have all been confirmed, as will be shown in chapter 5.

[23] This raises the problem of how the additional infrastructure for synchronization in the I-system is actually implemented in the brain. I will return to this issue in chapter 6.

4.2.2 Determinism and Neural Implementation

The algorithms described earlier for computing models for 'after' and 'before' sentences are, in a strict formal sense, deterministic. If states consequent on e_1 can be derived, and it can be checked whether these conflict with the conditions for e_2, then the algorithm always returns a model for a 'before' sentence, which either includes e_2 or not; and if no consequent states are computed, then there is no conflict with conditions on e_2, which can be added to the model. Processing here would remain fully deterministic even if we assumed that the computation or retrieval of consequent states or consistency checking may not terminate.[24] That may not introduce uncertainty into the algorithm but would only make the resulting models partial. When considered purely as formal objects, algorithms may or may not be deterministic, typically depending on their logical structure. In a classic paper, for example, Hobbs and Shieber (1987) present and compare a deterministic and a nondeterministic variants of an algorithm that enumerates quantifier scopings. The goal is to generate all and only the valid readings, not just those preferred by comprehenders ("heuristically primary" readings), or all possible readings (a sentence with n quantified NPs has $n!$ different scopings). The nondeterministic algorithm returns a single scoping—likely a different one each time the algorithm is run. The deterministic algorithm computes the same functions iteratively, and it returns the full set of valid scopings.

From a cognitive neuroscience stance, this notion of determinism as a formal property of algorithms is empirically vacuous, however. Whether a system that instantiates given (deterministic) algorithms actually behaves deterministically depends on details of bottom-level neuronal implementation. If one formulates this as a formal problem, in the framework of Turing machines (TMs), one can prove that, for every nondeterministic TM, there is an equivalent deterministic TM that carries out ("simulates") the same computations. However, this comes with costs, in terms of computational space (storage) and time for deterministic TMs (Savitch, 1970; Book, 1974; Paul et al., 1983; Cook and Nguyen, 2014). One example is pronoun resolution in DRT. Recall how the minidiscourse (2) ('A dog barks. It is black') is processed: first, a DRS for the first sentence is set up, introducing a discourse referent (**x**) for 'a dog;' next, a DRS for the second sentence is built up, with a discourse referent (**y**) for 'it;' finally, the two DRSs are merged, and the two discourse referents are unified ($x = y$). But how does the system decide which variables should be unified? While discourse imposes

[24] There is no reason why the derivation of the effects and noneffects of an event should be a finite process. In principle, these form open-ended sets as in the "frame problem" (McCarthy and Hayes, 1969; McDermott, 1987). The consequences of nonterminating computation in a logical calculus of events are discussed by van Lambalgen and Hamm (2004b).

constraints on the resolution of pronouns, it does *not* in general provide unique solutions for unification. Determinism may be easily reintroduced, if pronouns are "coindexed" with their antecedents in the syntax (Groenendijk and Stokhof, 1991; Chierchia, 1992, 1995; Muskens, 1996, 2011). As the general results for TMs and the DRT example show, nondeterministic computations may often be given a fully deterministic treatment.[25] What allows one to determine whether a neural process is deterministic is ultimately its actual implementation, which can only be investigated experimentally or indirectly via simulations.

At the neural level of explanation, nondeterminism is forced on deterministic and nondeterministic algorithms alike by the fact that the brain is a (stochastic) machine whose behavior lies between chaos (or sensitivity to initial conditions) and perfect robustness (Mannino and Bressler, 2015). In general, the source of indeterminacy is not just noise in the system (e.g., from the input) but a number of endogenous variables that may affect the execution of most algorithms in the R- and I-systems, particularly *local variables* (discourse), *global variables* (the recent history of computation in the system), *dynamic variables* (the effects of learning on stored data), and *concurrent processes*, that operate on some of the same representations as the R- and I-systems (e.g., perception and reasoning). Under the influence of those variables, semantic algorithms effectively become nondeterministic (e.g., similar inputs may not always lead to the same outputs). For example, the algorithm for processing 'before' sentences might well result in different minimal models on successive executions (e.g., in which the article was and was not submitted, respectively), if knowledge that the journal's policy change may prevent a submission is acquired between the two executions. The nonlinguistic context may provide new referents through perception, memory, or inference in addition to those mentioned in the discourse, thereby increasing indeterminacy in interpreting DPs and NPs.

The algorithms for binding and interpreting representations (of linguistic and other expressions) implemented in the R-system and I-system, respectively, are quite unlike the kind of context-, knowledge-, and path-independent algorithms associated with, say, arithmetic operations, yielding the same output every time they are executed. Few cognitive processes in the brain can be implemented in, or approximated by, (serial) computations in a deterministic TM (Sigman and Dehaene, 2005; Sackur and Dehaene, 2009; Zylberberg et al., 2011); semantics is *not* one of them. Consequently, it is difficult to produce algorithmic analyses that have *predictive* power and to construct experiments in which such analyses may be tested while controlling for the effects of exogenous variables.

[25] Notice how this parallels the empirical vacuity of the principle of compositionality, as discussed here and in chapter 1 (Janssen, 1986; Zadrozny, 1994; Baggio et al., 2012b).

4.2.3 Completeness, Underspecification, and Iconicity of Models

If the goal of the I-system is to construct a *complete* minimal model, the output of (nondeterministic) algorithms could still be an *underspecified* representation of discourse (Sanford and Sturt, 2002). A broad range of phenomena is brought under the label of "underspecification." Despite the superficial diversity, these phenomena point to a common insight: it is often *impossible* or *unnecessary* to derive a fully specified grammatical analysis or discourse model.[26] What, then, is the status of computational theory, if it may be unnecessary or impossible for the I-system to attain its main computational goal? A discussion of the sources and nature of semantic underspecification will help clarify this question.

Underspecification has its roots in the R-system. It is built into the structures that the R-system computes with. Consider the relational vagueness inherent in nominal compounds. In 'university student,' 'university office,' and 'university computer,' the conceptual relations between the constituents are very different: the student is enrolled at the university; the office could be part of the university premises (concrete) or of an administrative section (abstract); and the computer belongs to the university. Differences in meaning could be even more elusive, as in 'Apple logo,' 'Apple computer,' and 'Apple employee' (Bunt, 2007). The effort required to clarify these semantic relations offline appears considerable. This makes it seem even less likely that we construct online (minimal) models, where these relations are made explicit in full detail. A minimal model where a relation $R(a,x)$ between the constant Apple (a) and a term to be unified with x (e.g., logo, computer, or employee) is satisfied would be "good enough" for the purpose of comprehension (Ferreira et al., 2002; Ferreira and Patson, 2007).[27] This kind of underspecified minimal model seems sufficient to represent what is understood by these expressions in many cases. The status of computational theory would not be diminished here, as the resulting model *is* in fact complete.

Another source of underspecification is polysemy. Retrieving (or accessing) different unrelated meanings of a word typically activates core cortical regions of the R-system, including LIFG (Hoenig and Scheef, 2005; Rodd et al., 2005; Davis et al., 2007; Zempleni et al., 2007). In chapter 2, it was hypothesized that these activations reflect processes supporting semantic binding, and not merely semantic selection (Hagoort et al., 2009). This argument can be expanded with

[26] In the literature, the term 'underspecification' is used with equal frequency to refer to a class of linguistic phenomena (e.g., vagueness, ambiguity, polysemy) and to a family of formal techniques for describing alternative readings of an expression using a single representation. See Bunt (2007), Egg (2010), and Traxler (2014) for reviews. For an early proposal, see Reyle (1993).

[27] One could consider 'Apple' as a complex relation, involving the set of all entities that are made by Apple, of individuals who work for Apple, or of anything that is in other ways related to Apple. However, underspecification would persist even in this case.

a discussion of the processing consequences of polysemy for the I-system, in particular for (minimal) models of discourse containing homonyms. Consider the following example:

(15) John made his way to the bank.

where 'bank' means either the terrain alongside a river or a financial institution. Psycholinguistic evidence here is mixed. Earlier studies, for example by Onifer and Swinney (1981) and Seidenberg et al. (1982), have suggested that in lexical priming tasks multiple meanings of a word are momentarily accessed, and that selection occurs only in certain cases (e.g., for noun-noun but not for noun-verb ambiguities). More recent studies show that specific meanings are not selected or computed until they are needed (Frisson, 2009). Frazier and Rayner (1990), using eye tracking, showed that, in the absence of disambiguating information, fixation times were longer for homonyms (e.g., a 'ball' meaning sphere, dance, or bullet) than for words with multiple related senses (e.g., a 'library' as a room or building, or a collection of books and periodicals). These data are consistent with the view that sense selection is minimized and occurs only when a choice is required in order for the language system to preserve discourse coherence.[28] Either because multiple word meanings are immediately accessed or retrieved, or because semantic specification is delayed, selection is neither automatic nor mandatory in discourse processing.

Polysemy may be reconciled with Johnson-Laird's theory of mental models, but these results challenge one of the core tenets of the theory, namely the idea that mental models are *iconic* (Johnson-Laird, 1983, 2004, 2005, 2010)—there is an isomorphism between the structure of a mental model and the structure of what it represents.[29] What is problematic is not the isomorphism per se. There may be degrees of iconicity, gaps, and empty placeholders that would preserve the desired one-to-one mapping. Rather, the difficulty is that it may not always be clear *what* mental models represent. For example, a mental model of

(16) The library is currently being rebuilt.

should be isomorphic either with the physical restoration of a building, or with the reorganization of an archive, but underspecification requires that the model capture the "gist" of both. This renders the model an abstract structure that no

[28] This type of result, if confirmed, may point to the existence of principles of "cost minimization" in semantic systems, balancing out the effects of asynchronous incrementality: not everything that can be computed is computed immediately.

[29] A notorious precursor of this view is Wittgenstein's "picture theory of meaning" (Wittgenstein, 1922). Johnson-Laird's historical account also mentions Köhler, Maxwell, and Peirce among the forerunners in the development of mental models (Johnson-Laird, 2004).

longer represents iconically. That is precisely what a minimal model would do. In a minimal model of (16), the library is represented as an *incremental object* whose stage of completion varies over time, as the consequence of the building activity, represented as ongoing. This kind of representation is compatible with both readings of (16), and it is not substantially modified by specifying whether it is a building or an archive that is being rebuilt. It is fair to say that a modicum of underspecification is built into most minimal models. For instance, the bank in (15) is represented as a spatial location or region that John has reached in a finite amount of time, and that is all the discourse asserts.[30]

There is full agreement between a computational analysis that posits minimal models as the goal of interpretation and algorithmic analyses that characterize interpretation as an incremental and nondeterministic process that can generate underspecified representations. Semantics and psychology (neuroscience) can therefore be reconciled (Hamm et al., 2006). But how are they related exactly? The competence hypothesis was originally formulated for accounts of syntactic parsing (Steedman, 1992; Shieber and Johnson, 1993), but it specifies possible relations between competence and processing that also apply to semantics:

(a) Strict competence (STC) demands that discourse processing algorithms are *determined by* the competence theory. In DRT, partial models are built up incrementally, taking into account contextual information, computation is deductive (Kohlhase, 2000; Kohlhase and Koller, 2003), and models are fully specified semantic structures.

(b) Strong competence (SGC) requires that computational analyses *control* processing, but underspecified meanings are possible. In the event calculus (van Lambalgen and Hamm, 2004b), a single (minimal) model of discourse is computed *deductively*, but that may be an underspecified representation that captures the "gist" of alternative interpretations.

(c) Weak competence (WKC) entails that discourse models are, in general, those sanctioned by the computational analyses, but the actual construction process is not typically deductive and follows abstract principles of cortical computation, that is, incrementality, immediacy, parallelism, (a)synchrony, nondeterminism, and weakly modular functional integration.

STC does not seem to be a viable model here, but it remains an open issue just how much of SGC can be retained (e.g., how much of cortical computation can be modeled as a deductive process).

[30] For additional examples of (minimal) models of discourses about events, see van Lambalgen and Hamm (2004b). The (minimal) models described in their book are neither iconic nor isomorphic with the actual events they represent.

4.2.4 Functional Integration, Functional Separability, and Modularity

One last issue, concerning processing principles and algorithms in the I-system, is the extent to which binding and interpretation can be regarded as functionally separable. The circular model adopted here entails a high degree of functional *integration* between forms of binding (R-system) and interpretation (I-system). Alternatively, a sequential scheme, where composition feeds into interpretation sentence-by-sentence, is aligned with complete systems separability. Here, the idea of modularity becomes relevant. Fodor (1983) listed and discussed the key properties characterizing a modular system: domain specificity, fast mandatory operation, limited central access to the representations computed by a module, informational encapsulation, shallow (e.g., uninterpreted) outputs, fixed neural architecture, distinctive breakdown patterns, ontogenetic sequencing and pace. Most relevant to interpretation is *informational encapsulation*. Fodor applies it to the process that yields a "structural description" of an input sentence (Fodor, 1983, p. 44), similar to the R-system's computation of type-to-token relations. If that process also comprises binding and, in parallel, checking of grammatical constraints (chapter 2), then one can understand that Fodor is arguing precisely *against* a top-down account of the relations between interpretation and binding, whereby general information available to the hearer is fed back from relatively high levels of representation to the process of structural analysis. The effects of discourse on lexical semantic activation (N400) and the constraints imposed by interpretations on phrase- or sentence-level structure (P600) are two examples. One can only agree with Fodor on a default position, in which binding is likely encapsulated from *some* types of information. Similarly, *some* algorithms must compute specific aspects of relational semantic and morphosyntactic structures in isolation from background information. The question is, does interpretation, in particular, affect these processes (including binding), and if so, how?

In a crucial and quite controversial passage of his argument for informational encapsulation, Fodor suggests that "feedback works only to the extent that the information which perception supplies is redundant" (Fodor, 1983, p. 67). He then proceeds to relate feedback to the *prediction* of aspects of the input, based on higher-level information. Fodor notes that a sentence such as

(17) I keep a giraffe in my pocket.

cannot be predicted from the context, "on even the most inflationary construal of the notion of context" (p. 67). Discourse is often unpredictable (Jackendoff, 2002), prediction may be unnecessary, and it is not an overt computational goal (chapter 1). Moreover, bottom-up processing occurs and occasionally suffices. But Fodor's reasoning appears flawed. One cannot argue against the contextual penetrability of language processing by using a case in which context is empty:

(17) is given as an isolated token. The possibility of bottom-up composition in the absence of context is not a good argument for informational encapsulation. Rather, Fodor should show that context has no effect *when present*. Everything that is known on semantic binding, much of which derives from N400 research, points in the opposite direction. Semantic binding is eminently sensitive to the discourse context and radically predictive or, more accurately, it builds heavily on the effects of lexical semantic preactivation (chapters 1–3) (Hagoort and van Berkum, 2007; Hagoort et al., 2009; Kutas and Federmeier, 2011; Kutas et al., 2011; van Berkum, 2012). The time course of the N400 is strong evidence that the interplay between context and lexical (semantic) processing is as fast as the latter may be. This makes it difficult to rescue encapsulation and modularity by arguing that interactions are mediated by central systems (e.g., via reasoning). Conversely, not all contextual information that may influence binding online is encoded lexically (e.g., gestures), which makes it hard to salvage encapsulation by assuming that interactions occur *within* the module.

Modularity and compositionality are closely related notions. One may argue that the computations performed by a modular system are those regimented by compositionality, if indeed the output of the module (the meaning of a complex expression) is a function of the information available to the module: syntax and the lexicon.[31] Alternatively, one can envisage a (serial) system of *two* modules: one that produces syntactic trees for sentences, which are then fed into another module, containing lexical meanings and rules of combination, corresponding to semantic binding and interpretation operations, finally yielding the meaning of the complex expression, that is, a *sentence* model.[32] Here, compositionality constrains the correspondence between syntactic and semantic combinatorics, that is, the traffic that may occur between the two modules. Regardless of one's choice of a specific modular architecture, however, any influence of the context (broadly construed) on composition or interpretation is assumed to occur *after* the module(s) has (or have) produced an output and to be mediated by "central systems." Again, N400 evidence for the effects of discourse context on lexical semantic processing (chapter 2) is arguably the strongest evidence available at present that the output of the I-system affects binding in the R-system, as in an incremental and interactive processing model.

---

[31] Jackendoff makes essentially the same connection: "The hypothesis of syntactically transparent semantic composition has the virtue of theoretical elegance and constraint. Its effect is to enable researchers to isolate the language capacity—including its contribution to semantics—from the rest of the mind, as befits a modular conception [...] However, it is only a hypothesis [...] whatever its a priori attractiveness, it cannot be sustained." (Jackendoff, 1997, p. 49)

[32] For a modular approach to meaning, see Borg (2004). For a discussion using philosophical and linguistic arguments as well as neural data, see Robbins (2007, 2013) and Cosentino et al. (2017).

4.3 Three Dimensions of Interpretation

We must distinguish between three types of I-system processes, corresponding to three "dimensions" of discourse interpretation: *referential processing* (e.g., assigning referents in the discourse model to tokens within the active relational structure, such as the referents of nouns and pronouns), *elaborative processing* (e.g., extending the discourse model to include pragmatic, figurative, and other aspects of meaning), and *inferential processing* (e.g., exploring the structure of the discourse model and fleshing out, at the level of the relational structure, the model's logical consequences, taking into account the context and background knowledge). This tripartite classification is somewhat arbitrary. Moreover, the boundaries between the three dimensions are largely theory dependent. Indeed, we assume that there are several open questions (e.g., whether pragmatics boils down to inference and whether figurative meaning is derived compositionally), whose answers may require a transfer of linguistic phenomena, spanning levels of analysis, in particular across the elaborative-inferential boundary. However, while these questions are being studied by linguists and philosophers, cognitive neuroscientists can explore tentative classifications to determine whether these mirror distinctions at the neural level of analysis. The claim here, which will be further substantiated in chapter 5, is that referential, elaborative, and inferential processing are different *modes of operation* of the I-system. They are predicted to leave traces in M/EEG signals and to result in specific fMRI activations, but they are not competing or mutually exclusive processes in the brain.

Predictions are straightforward for most basic cases of referential processing. First, representations of referents ("denotations") are computed in the I-system *after* representations of types ("senses") and tokens are activated and bound by the R-system. Consequently, one should expect referential processing to inflect components of the ERP signal that *follow* the N400 in time. Second, these ERP responses may not be phasic, but instead *tonic*: they would not set on and off at fixed times, as the N400 does, but would show greater variability in onset times or duration. Punctuated processing (e.g., clause-by-clause incrementality) and nondeterminism in binding tokens to referents in the active discourse model are among the causes of slower ERP effects. Third, one should expect *independent activations* in brain regions recruited by various forms of referential processing relative to the R-system networks (LIFG and pMSTG) in experiments in which referential and relational processing are manipulated independently. M/EEG or fMRI data supporting these predictions would suffice to establish empirically a functional distinction between the R-system and the I-system in the brain, even if functional associations (as opposed to dissociations) were observed between R-system processing and instances of elaborative or inferential processing.

One would similarly expect a degree of functional independence between the networks that support forms of inference, such as deduction, and the R-system. Stenning and van Lambalgen (2008) distinguish reasoning *to* an interpretation (imposing a particular logical form among several possible on the input, so that a model may be computed) from reasoning *from* an interpretation (drawing and withdrawing inferences from the model). Here, one should exercise some care, because what is viewed as reasoning *to* or reasoning *from* in each case depends on one's theoretical choices. One could analyze relational semantic processing as inference (recall the discussion of predictive inference in chapter 1), treating much of what happens within the R-system as "reasoning" to an interpretation. At least some (but formally all) algebraic operations on semantic vectors, in the service of top-down and bottom-up binding, may be captured by a combination of different forms of deduction. A more interesting possibility here is that both reasoning *to* and reasoning *from* an interpretation are carried out by and within the I-system, while R-system processing is (largely) noninferential. Indeed, the relational structures that result from binding are typically much richer than the logical forms that a semantic theory would assign to the relevant strings. It then falls to the I-system to *select* what exactly, in a relational structure, deserves to be interpreted and to become part of the model (reasoning *to*), and accordingly to compute a model, mapping selected tokens from the relational structure onto the reference structure. The internally generated logical forms that mediate this process might be different from the logical forms derived from (compositional) analyses of input strings (e.g., as complement coercion demonstrates). There is very little experimental evidence that can be brought to bear on this hypothesis. However, one may expect that LIFG is involved (since it has been shown to be activated when selection demands increase) with working memory or attention systems in the parietal lobes. Chapter 5 includes a discussion of fMRI evidence that the left parietal cortex plays a role in generating multiple interpretations of a given token structure (e.g., for the material conditional, $p \rightarrow q$). Experiments will be presented that have succeeded in identifying ERP signatures and brain networks underlying reasoning *from* an interpretation.

Elaborative processing demands more careful, in-depth analysis, particularly of one of its most ubiquitous manifestations: metaphor. Elaborative processing is variously defined in psycholinguistics. McKoon and Ratcliff (1989) describe elaboration as the process of adding meaning to asserted content,[33] which takes

[33] In the literature, one may find broader views of elaboration, including logical inference (Brooke, 1995), and narrower views, in which elaboration would merely provide rhetorical functions. These accounts, however, seem less standard and theoretically less productive than the model by McKoon and Ratcliff (1989, 1992) discussed here.

into account the goals of the comprehender and builds on a small set of (largely automatic) inferences on discourse content and background knowledge—what constitutes a minimal model. These inferences are required to render discourse coherent (McKoon and Ratcliff, 1992). This account aligns with the distinction proposed here between inferential and elaborative processing, but importantly not all forms of elaboration (that is, interpretive processes that do not belong to either the referential or the logical-deductive types) are of an inferential nature. However, they could be formally analyzed as inferences, as is the case for much R-system processing. Metaphor is a case in point. Consider for example 'John is a fox.' Here, one has to derive the intended meaning (that John is sly) from a strictly false premise (that John is a fox). That may be an insuperable challenge for classical deductive inference, but one could restore the possibility of a fully inferential account of metaphor via modification of one's theory of predication. One need not assume that 'fox' can only refer to the animal category. Metaphor vehicles, such as 'fox,' could be thought of as having *dual reference*: foxes and the class of all individuals, which may not include actual foxes, with particular characteristics (e.g., cunning). In this view, a metaphor is *literally true*. It then remains to be seen whether its meaning may be derived inferentially or whether processing costs (e.g., sense selection) are instead transferred to the R-system. Cases in which dual reference is explicit can be handled by the R-system.

(18) Cambodia was Vietnam's Vietnam.

Understanding this sentence requires that two different (token) representations of the expression 'Vietnam' be generated and maintained (by LIFG, according to the cycle model; chapter 3), linked to different stored (type) representations of Vietnam as the country in Southeast Asia, in one case, and of the category of prolonged and disastrous armed conflicts exemplified by the war in Vietnam, in the other case. Metaphor (and metaphor theory) will be further discussed later. A provisory conclusion is that elaborative processing, as typified by metaphor, is not a unitary phenomenon. Occasionally, it straddles the boundary between relational and interpretive processing.

Metaphor reminds us moreover that several independent "layers" of meaning may be identified in what one experiences as the intended speaker's message. It is the task of a computational-level analysis of metaphor to provide an adequate representation of each layer and of relations between layers. Suppose that (19) is uttered sarcastically

(19) Max is turbocharged.

and consider what goes into a model (Berg, 1988). The literal meaning, derived from the logical form or propositional representation of (19), cannot lead to the correct interpretation and neither can the (metaphoric) meaning that Max is full

of energy. Understanding the intention of the speaker here leads to the *negation of the metaphoric meaning* (i.e., that Max is in fact lazy). So, not only are there different layers of meaning, but also possibly frictions, occasionally even overt contradictions, between layers. Such tensions between literal, metaphoric, and pragmatic meaning are understood differently in different theories of metaphor (Lyon, 2000). Different theories often have different processing consequences. Four types of analyses of metaphor have been put forward in logic, linguistics, and philosophy: *semantic*, *logical*, *pragmatic*, and *hermeneutic*.[34]

Semantic analyses descend from traditional theories of metaphor as simile or comparison, in which 'John is a fox' means the same as 'John is like a fox,' but can often depart significantly from it. Black (1954, 1979) developed a semantic analysis of metaphor that takes issue with the notion that metaphor involves the mere substitution of literal meaning—"saying one thing and meaning another." The latter notion presupposes that a metaphoric expression M can be translated into a literal expression L, without loss of meaning. In figurative language use, the speaker or author typically provides not the intended meaning, m, but some function of it, $f(m)$. The listener or reader then applies the inverse $f^{-1}(f(m))$, which should restore m. The interpretation function f^{-1} varies, also depending on the relevant trope. In some forms of irony and sarcasm, f^{-1} may reverse the effect of f. In metaphor, it establishes a comparison or a mapping between two terms. That is what the semantic system as a whole is required to do: compute a *range of values* for $f^{-1}(f(m))$. The most pressing question here is whether this computation occurs in the R-system or I-system. It is theoretically possible to see f^{-1} as a purely interpretive transform that need not alter lexical meanings. Thus, in 'John is a fox,' the word 'fox' would maintain its literal meaning, and the sentence would be interpreted as though it had the logical form of a simile. Black aims to argue against precisely this type of account. He distinguishes the *frame* (what must be interpreted literally) from the *focus* (the metaphor carrier) of a metaphoric expression, and he argues that metaphoric meaning arises from the interplay of frame and focus. The literal sense of the focus ('fox') should be shifted or restricted, so that only some properties (e.g., cunning) may be related to the frame ('John'). This process of sense adjustment is more relational than interpretive, as it implies selecting and binding lexical features—a process that precedes and enables model construction. However, Black's theory entails that sense adjustment often requires taking into account the speaker's intentions in context, much as in the pragmatic accounts presented below.

[34] These umbrella terms are only useful to cluster together a number of accounts that share certain formal features or a general outlook, but they should not be taken to imply that theoretical diversity within each cluster can or should be reduced or eliminated.

A type of semantic analysis that extends and modifies Black's sense-shifting theory is based on the idea that the semantic relations highlighted by metaphors are not between individual concepts (e.g., frame and focus) but between whole *experiential domains* (Lakoff and Johnson, 1980; Lakoff, 2008).[35] Ultimately, it is not language but thought that is inherently metaphorical. For example, the ideas that 'argument is war,' or that 'time is money,' involve networks of related concepts, words, and phrases: 'to win an argument,' 'to defeat a detractor,' 'to spend time,' 'to invest time.' These stored relational structures are the products of cultural processes of modification and transmission of information, drawing on different kinds of experience. Less often (e.g., in literature) they result from an individual creative act. In other analyses, it is the *specificity* of the semantic relations involved in creating and processing metaphors that matters. Metaphor would, therefore, rely not so much on a generic mapping between two domains, but on *blending* specific elements from those domains (Fauconnier and Turner, 1998; Coulson and van Petten, 2002), as in

(20) That surgeon is a butcher.

What results from binding here is a conceptual structure in which the agent and goal are those typical of surgery, while the means and manner are characteristic of butchery. Blending may address what Black regarded as the main weakness of his analysis: that the process of selection and extension of semantic features, transferred from the focus to the frame, is often quite hard to pin down formally (Black, 1979). A more systematic effort to capture in formal terms the relations between the elements of a metaphor is made by logical accounts.

One aim of logical analyses is to explain the fact that a metaphor may be true or false in context. That is, we may view metaphor comprehension as involving the *construction of a model of discourse making the metaphor true*. Metaphors, too, would abide by Miller's law. In logical accounts, a metaphor is assumed to involve the violation of selectional restrictions, such that predicates are applied to arguments of an "incorrect" type, category, or sort. As was observed earlier, complement coercion has been analyzed as a response to a type mismatch (e.g., between the entity referred to by 'the book' and the event required by 'began'). Likewise, metaphor may be seen as a case of *sortal incorrectness* (Thomasson, 1972; Bergmann, 1977). For example, (19) is sortally incorrect. Human beings such as Max are not the sort of entity that can be 'turbocharged.' Only engines are. One could devise a model-theoretic semantics (Guenthner, 1975; van Dijk, 1975; Bergmann, 1979; Lappin, 1981; Steinhart, 2001) in which the customary domains of interpretation of predicates would be modified (Nunberg, 1995), so

[35] See also Tourangeau and Sternberg (1982) for a related "domains-interaction" theory.

that one may truthfully (literally) utter (19), 'She's an encyclopedia,' or 'He's a bear.' In this way, predicate extensions are expanded or reduced, but they may also undergo selection processes at the level of the relevant reference structure. Semantic features, associated to lexical predicates, that are normally applied to more restricted regions of logical space (e.g., to engines or books), are dropped, so that the product can be applied to entities of a different sort (van Dijk, 1975). The relationships between sortal incorrectness and metaphoricity are complex. Not all sortally incorrect phrases may be interpreted metaphorically, and not all metaphors are sortally incorrect (Lappin, 1981); for example, (20) is not. Thus, logical analyses can provide only an incomplete account of metaphor. Yet, they are relevant for our purposes, as they locate the sources of metaphoric effects in the definition and computation of interpretation functions, therefore effectively in the I-system. It is possible to develop explicit formalizations of the reference structures involved in certain metaphors; for example, in the transfer of specific geometric properties across (potentially isomorphic) domains (e.g., in 'peak of a mountain' and 'peak of a career') (Gärdenfors, 2000, pp. 176–187).[36]

Some of the processes involved in metaphor comprehension, as envisaged by semantic and logical analyses, may be captured algorithmically. The model by Kintsch (2000, 2008), for example, addresses two characteristics of predication in metaphor: (a) the fact that understanding metaphors involves selection of the right features from sets made available by the metaphor's carrier, like energetic in (19); (b) the fact that metaphorical predication is not reversible; for example, 'My lawyer is a shark' is acceptable as a possible metaphor, while 'My shark is a lawyer' is not.[37] Kintsch applies Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997), where, as in the distributional theories discussed in chapter 1, meanings are formalized as vectors or points in a high-dimensional semantic space. The similarity between the meanings of two words is then computed via the cosine of the angle between the two vectors that represent them (Erk, 2012). This operation on vectors is appropriately order insensitive, but binding may be exquisitely order *sensitive*, as is shown by 'My shark is a lawyer.' The problem with commutative operations, designed to capture meaning composition, is that they are all order insensitive and fail to account for the fact that different orders of predication produce different meanings. The algebraic operations discussed in chapters 1 and 3, in particular, would still require a constraint, implemented at the algorithmic level, that two alternative orders of predication could lead to

[36] For model-theoretic and algorithmic analyses of domain transfer and equivalent notions, see the work by Indurkhya (1986, 1987, 2013).

[37] There are exceptions. For example, 'This butcher is a surgeon' appears to allow for a metaphoric interpretation that may be fully accounted for in terms of blending, as is the case for (20).

two different relational structures, of which only one allows for a metaphorical interpretation. Kintsch addresses this problem with the following algorithm:

(i) take all semantic features related to the relevant predicate ('shark');
(ii) select those features of the predicate that can be related to the argument ('lawyer;' e.g., viciousness) consistent with stored knowledge;
(iii) compute the semantics of the metaphoric expression via the centroid[38] of the argument vector and of the vector resulting from (ii).

The sequence of steps here explains why, for example, applying (i) to 'lawyer' and (ii) to 'shark' would produce a different (possibly anomalous) result. Thus, the asymmetric nature of metaphoric predication flows smoothly from this type of analysis. Here, the issue is whether this may be seen as a *universal* metaphor processing algorithm, so that for conventional metaphors one activates a stored representation corresponding to the vector that results from (iii), and for novel metaphors one carries out the computation from scratch.

Interpretation is given a more prominent role relative to R-system processing in pragmatic and hermeneutic accounts. In a now-classic analysis of metaphor, Searle (1978, 1989) argued that metaphoric meaning should be regarded as part of speaker meaning (Grice, 1957). We say something, and *what* we say means literally, but what *we* (speakers) mean by it may be different (e.g., a metaphoric meaning). The case for a pragmatic treatment of metaphor appears compelling, as metaphors violate all of Grice's conversation maxims (Grice, 1975): Quality (metaphors are false), Quantity (metaphors seem to convey little information), Manner (at least novel metaphors appear obscure or ambiguous), and Relation (they are not relevant, since they literally speak of something else) (Eco, 1986). Therefore, metaphors can be expected to trigger one or more implicatures. This pragmatic process would allow the listener to recover meaning as was intended by the speaker, under certain assumptions. This pragmatic (inferential) account is effective in at least two cases. First, it functions better with novel metaphors, whose meaning must indeed be recovered, while its motivation appears weaker with conventional and dead metaphors, which may not trigger implicatures and whose meaning, instead, may simply be retrieved. Second, it works well only if metaphoric meaning is identical to speaker meaning, and that is not always the case. In the sarcastic comment (19), speaker meaning would imply that Max is lazy. Consequently, it cannot entail its opposite, the default metaphoric reading that Max is energetic. Contra Searle, Berg (1988) and others have claimed that metaphoric meaning is, like sarcasm, the result of the application of one among several operators that would jointly turn literal meaning into speaker meaning.

[38] Alternative algorithms may be formulated using different composition operations.

These operators may be applied sequentially—*interpretations are derived from previous interpretations*. The idea that Max is lazy is derived from the idea that Max is full of energy (via the sarcasm operator), which is derived from the idea that Max is turbocharged (metaphor operator). The sequential character of this theory may turn out to be implausible from a processing perspective. However, Berg's analysis predicts that metaphors are processed by the I-system, in ways that may differ from standard pragmatic processing.

There are several accounts that do not view metaphor in terms of implicature or speaker meaning, but rather emphasize reasoning *to* an interpretation in the discourse context in which the metaphor appears. These may be considered as pragmatic accounts, their logical flavor notwithstanding. Asher and Lascarides (1995, 2001, 2003) have constructed a version of DRT in which constraints on senses and on their metaphoric extensions are encoded lexically. In this theory, interpretations, based on such constraints, always connect to the wider context. The assumption is that metaphoric senses, much like word meaning in general, are underspecified, and specification is constrained by the discourse. Metaphor is seen as requiring minimal changes to stored meanings, and hence predicates may be applied in partial violation of lexical constraints. Consider

(21) John is a rock.

One could derive the "best metaphoric interpretation," as Asher and Lascarides call it, by restricting the sense of the predicate 'is a rock'[39] to anything heavy, solid, and hard to move, so that predication, even applied to a human argument, would succeed. The notion that predicates have to be modified before they can be applied in a metaphoric context is common to semantic and logical accounts and to Kintsch's algorithm. What is characteristic of the analysis by Asher and Lascarides is that the *resulting* metaphoric meanings are just as underspecified and defeasible as any lexical meaning. Suppose that (21) is followed by

(22) But (compared to John) Sam is a pebble.

Now (21) acquires a more specific meaning, in light of (22), that does not seem to involve solidity as a feature relevant to the metaphoric meaning: size is what matters here. Shifting senses is not strictly a relational computation, but rather a process of nonmonotonic inference that takes discourse context into account. The prediction then is that metaphor comprehension recruits the I-system and, in particular, cortical networks that are involved in deduction or other forms of (defeasible) reasoning.

[39] Technically, this is an instance of predicate circumscription (McCarthy and Hayes, 1969). Like closed-world reasoning, circumscription is nonmonotonic.

In a more radical proposal, Davidson (1978) argued that metaphors can mean nothing less (or more) than what the carrier expression means literally, and that the "metaphoric effect" results from the *work of interpretation and imagination* by the reader or listener. That is the basic idea of hermeneutic accounts.[40] For Davidson, in particular, there is no such thing as metaphoric meaning: nothing that can be carried by the expression, as in semantic and logical accounts, or by the speaker, as in pragmatic accounts. If metaphors have any effects on thought and the imagination, it is these internally generated representations, rather than the metaphoric expression itself, that may be given a meaning and a truth value. Recovering figurative content from metaphors or other tropes is a constructive, open-ended hermeneutic exercise. The centrality of interpretation in metaphor comprehension has been emphasized by Ricoeur (1974, 1977) and Eco (1984, 1986), among others. More specifically, Eco has defended the view that lexical knowledge (dictionary) typically underdetermines metaphoric meaning, which must be further specified through the active involvement of general knowledge. From the encyclopedia, one would extract the "semiotic features" to be applied to the metaphor's argument. Such features are not necessarily semantic, that is, perceptual or conceptual, but may be drawn from a broader "cultural database," which renders interpretation noncomputable algorithmically.[41] Eco's proposal contrasts most sharply with semantic and logical accounts. Ricoeur's addresses instead the assumptions of pragmatic accounts. His analysis raises the problem of the interpretation of *literary texts*, in particular. Because one cannot engage in a conversation with a work of literature, pragmatic principles of cooperation must be replaced by specific hermeneutic principles. *Hermeneutic asymmetry*, between a static text and a dynamic interpreter, displaces *pragmatic symmetry*, between a dynamic sender and a dynamic receiver.[42] Ricoeur emphasizes the open-ended nature of interpretation. To interpret means to identify and clarify

[40] At the outset of his paper, Davidson writes that the metaphoric effect results from the "interplay of inventive construction and inventive construal;" however, his account effectively emphasizes the interpreter's side. I will discuss metaphor production, and semantic production more generally, in chapters 7–9.

[41] My understanding of Eco's thesis here is that noncomputability should be taken as a "normative constraint," and not as an empirical claim. For example, he writes: "That metaphor is 'good' which does not allow the work of interpretation to grind to a halt [...] but which permits inspections that are diverse, complementary, and contradictory." (Eco, 1986, p. 120)

[42] A case can be made that experimental situations involve a similar kind of asymmetry, where the source of experimental stimuli is fixed, hidden, and shielded from interactions with the participant. This detracts from the ecological validity of many traditional experimental paradigms in cognitive neuroscience and psycholinguistics (Willems, 2015), but it also underscores the relevance and the potential of a hermeneutic approach to discourse processing in the laboratory. I return to the more specific issue of (a)symmetry between senders and receivers in chapter 7.

a domain of *references*—that which the discourse is about—rather than senses. This view is highly relevant for (but not limited to) the interpretation of literary metaphors, which typically requires finding suitable referents inside or outside the text for several expressions in discourse, not only for the metaphor's focus or vehicle (Reinhart, 1976). For example, take the line from "The Love Song of J. Alfred Prufrock" by T. S. Eliot

(23) I have seen them riding seaward on the waves.

where 'them' refers to the mermaids previously mentioned in the poem. If the figurative analysis process was restricted to 'riding,' the resulting interpretation would be partial. The mermaids would be depicted as floating on water, as may be consistent with the image Eliot meant to convey. But nothing would happen to 'waves.' There would be no representation of horses in relation to the waves. Here, a semantic "double perception" process is needed (Reinhart, 1976), that relates (a) 'riding' to floating, sharing the core dynamic feature of the reference structure (being on top of something in motion), (b) 'waves' to horses, namely the entities in the reference structure that display similar patterns of undulatory motion, and (c) both 'riding' and 'waves' to the cultural reference frame within which a correlation between these two domains is preferred to other mappings. Hence, metaphor interpretation includes domain transfer but cannot be reduced to it. Understanding literary metaphors requires a hermeneutic exercise, which is amenable only in part (e.g., in some constituent operations) to a relational or a logico-inferential analysis. The hermeneutic account would then predict that figurative language comprehension is a temporally extended process, engaging an array of cortical networks (e.g., long-term memory, inference, and imagery) where activity may vary as a function of the number of expressions in discourse undergoing figurative interpretation and of the cultural background knowledge and hermeneutic capacity of the reader.