ABSTRACT

Title of Document: ON UTTERANCE INTERPRETATION AND METALINGUISTIC-SEMANTIC COMPETENCE

Kent W. Erickson, PhD, 2012

Directed By: Dr. Paul M. Pietroski, Department of Philosophy and Department of Linguistics

This study explores the role of what I call *metalinguistic-semantic competence* (MSC) in the processes of utterance interpretation, and in some cases expression interpretation. MSC is so-called because it is grounded in a speaker's explicit knowledge of (or beliefs about) the lexically-encoded meanings of individual words. More specifically, MSC derives, in part, from having concepts *of words*—or *concepts*$_W$ as I distinguish them—whose representational *contents*, I propose, are corresponding items in a speaker's mental lexicon. The leading idea is that once acquired speakers *use* their concepts$_W$ to form explicit beliefs about the meanings of words in terms of which extralinguistic concepts those words can (and cannot) coherently be used to express in ordinary conversational situations as constrained by their linguistically-encoded meanings. Or to put the claim differently, I argue that a speaker's explicit *conception* of word-meanings is a direct conscious reflection of his/her tacit understanding of the various ways in which lexical meanings guide and constrain without fully determining what their host words can (and cannot) be *used/uttered* to talk about in ordinary discourse. Such metalinguistic knowledge, I contend, quite often plays crucial role in our ability to correctly interpret what other speakers say. The first part of this work details the cognitive mechanisms underlying MSC against the backdrop of a Chomskyan framework for natural language and a Fodorian theory of concepts and their representational contents. The second part explores three ways that MSC might contribute to what I call a speaker's core linguistic-semantic competence. Specifically, I argue that MSC can help explain (i) how competent speakers acquire conceptually underspecified words with their lexical meanings, (ii) the contextual disambiguation of inherently polysemous words, and (iii) the informativeness of true natural language identity statements involving coreferential proper names. The philosophically relevant conclusion is that if any of these proposals pan out then MSC constitutes a proper *explanandum* of semantic theory, and hence any complete/adequate theory of semantic competence.

ON UTTERANCE INTERPRETATION AND METALINGUISTIC-SEMANTIC COMPETENCE

By

Kent W. Erickson

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
PhD, 2012

Advisory Committee:
Dr. Paul M. Pietroski, Chair
Dr. Peter Carruthers
Dr. Susan Dwyer
Dr. Erin Eaker

Acknowledgements

I owe a special debt of gratitude to my committee members, and in particular Paul Pietroski for his time, patience and direction throughout this project. Special thanks also goes out to Erin Eaker for insightful discussion and helpful comments on early draft chapters. I would also like to extend my gratitude to Mitch Green, who as my first real mentor provided both the basic tools and inspiration to pursue my ambitions in philosophy. In addition, I would like to recognize classmates Vincent Picciuto, Dimiter Kirilov, Yu Izumi, Harjeet Parmar, and Xuan Wang for their helpful discussion and general camaraderie over the past several years. Finally, I wish to acknowledge the unyielding support of family and friends in the pursuit and attainment of this goal.

Table of Contents

1. Introduction

1.1. On utterance interpretation and semantic competence

Consider the nonce sentence in (1), excerpted from Lewis Carroll's *Through the Looking Glass* (1871):

(1) All *mimsy* were the *borogoves*

In tribute to Carroll, (1) is an example of what language researchers call "Jabberwocky" sentences. Such constructions are often used in developmental and neuroimaging studies to diagnose how ordinary yet fully competent speakers parse and interpret various sentence-types in relative isolation from knowledge of the precise meanings/denotations of their open-class lexical constituents. Most relevant to my purposes here, such examples are occasionally invoked by philosophers of language to demonstrate our capacity to assign a coarse layer of meaning to such sentence-types based solely on our understanding of their grammatical structure and/or logical form.

Adapting an idea first advanced by Gilbert Harmon (1974), James Higginbotham (1988) has argued, for instance, that while we may be ignorant of the intended denotations/extensions of the nonce words 'mimsy' and 'borogoves', any competent speaker of English knows—if only tacitly—that (1) above has the same basic "logical skeleton" (i.e., grammatical structure) as a more familiar sentence such as (2):

(2) The plates were all broken

More specifically, the empirical evidence suggests that even very young native English speakers will know that 'borogoves' is a plural count noun that designates some class of things, or other, and that 'mimsy' is an ordinary adjective/predicate that describes some state or property of 'borogoves' (whatever those "things" happen to be). As importantly,

knowledge of the logical skeleton of (1) provides young language learners with tacit clues about what these words *cannot* mean. Higginbotham (*ibid.*: 166) notes, for instance, that "the child knows that the word 'mimsy' could not mean *tree*, and 'borogove' could not mean *run*." By hypothesis, this competence is adducible to what the child antecedently knows, which is (i) the formal grammatical structure of (1), (ii) the semantic principles by which sentences of this form are interpreted, along with (iii) the meanings of its other lexical constituents.[1]

The relevant point, which ultimately owes to Noam Chomsky, is that our tacit knowledge of the various ways in which the purely formal properties of linguistic expressions constrain without determining what those expressions can and cannot be used/uttered to express in ordinary discourse seemingly involves a fundamentally different aspect of linguistic competence as required to learn/understand the norm-governed denotations of their lexical constituents, and hence the thoughts/propositions (or truth conditions) that those expressions are customarily used/uttered to express.

More generally, utterance interpretation is commonly thought to depend on at least three distinct aspects of semantic competence: what I will call *linguistic-semantic competence* (LSC), *conceptual-semantic competence* (CSC), and *pragmatic-semantic competence* (PSC). First, I take LSC to be roughly what Chomskyans refer to simply as *linguistic competence*—competence grounded in a largely *tacit* (i.e., sub-personal, sub-doxastic) representational knowledge of the purely formal semantic properties of words together with the rules of grammar that collectively underwrite a speaker's ability to

[1] As Lust (2006: 185) explains, "Functional categories provide a critical mechanism by which we map the speech stream to the secret skeleton, just as we do for "Jabberwocky"… Children must link content words to the skeleton… When words are linked to the skeleton, children can label phrases; e.g., nouns will head noun phrases, verbs will head verb phrases.

generate and understand any of the infinitely many generable (i.e., possible) expressions of his/her native language. In terms of comprehension, LSC is manifest in a rather vague recognitional awareness of the various ways in which the meanings of expressions so-generated guide and constrain without fully determining what they can and cannot coherently be used/uttered to denote, describe, refer to, or otherwise talk about in ordinary discourse. For instance, LSC is posited to explain how it is that competent speakers of English—again even very young ones—seem to tacitly know that a nonce word such as 'borogoves' in the context of (1) above does not and indeed cannot designate, say, a *substance* such as water, or an *event* such as running.

CSC, by contrast, relates to a speaker's conceptualized world-knowledge of *what* his/her native language expressions are customarily used to talk about (i.e., factual knowledge about individuals, properties, relations, events, states of affairs, etc.). At the lexical level, we might think of CSC as subsuming a speaker's norm-governed ability to correctly align her words with the *concepts* that others in her local linguistic community customarily use those expressions to *express*. Lastly, as most theorists agree what I am calling PSC, or pragmatic-semantic competence, involves the Gricean ability to correctly infer from what is "literally" said/asserted by means of a linguistic utterance the thought/proposition that its utterer *intended to convey* or otherwise *communicate* in virtue of having said what she did relative to the context in which the utterance was produced.

So distinguished, the aim of this study is to demonstrate that the three aspects of semantic competence just outlined, either alone or in tandem, fail to exhaust one's core semantic competence—i.e., that body of linguistic information that one must know, or cognize, or otherwise *mentally represent* in order to be a semantically competent

language user. Specifically, I argue that whatever *else* semantic competence consists in, it depends, in addition, on what I call *metalinguistic-semantic competence* (MSC). MSC is so-called because it is grounded in a speaker's *explicit* (i.e., *consciously accessible*) conceptual knowledge of or beliefs about the purely formal, *intrinsic* semantic properties of linguistic expressions, and in particular the *meanings* of individual words.

Importantly, MSC owes neither directly to our tacit knowledge of language/meaning (LSC), nor to our conceptualized world-knowledge of the things we ordinarily use expressions of natural language to talk about (CSC), nor to our pragmatic conversational knowledge of a speaker's communicative intentions (PSC). Rather, MSC derives from having concepts *of words*—or $concepts_W$ as I shall distinguish them—whose representational *contents*, I propose, are corresponding items/entries in a speaker's mental lexicon. The leading idea is that once acquired speakers use $concepts_W$ to form explicit beliefs about the meanings of words (or lexical items) in terms of the constraints that their linguistically-encoded semantic properties impose on which *extralinguistic concepts* those words can (and cannot) coherently be used to express in ordinary conversational situations. Or to put the claim differently, what I will call a speaker's explicit *conception* of a word's meaning is by hypothesis a direct conscious reflection of his/her tacit understanding of the various ways in which its linguistically-encoded semantic properties guide and constrain without fully determining what that word can and cannot be *used/uttered* to denote, describe, refer to, or otherwise talk about in ordinary discourse.

Unlike most other discussions of semantic competence, my overarching thesis is that in certain cases MSC plays a *constitutive* role in utterance interpretation, and specifically with respect to the interpretation of expressions containing proper names. In

consequence, I contend that certain aspects of MSC should be treated as an integral component of a speaker's core linguistic competence, and more specifically what I am calling his/her *linguistic-semantic competence* (LSC). Moreover, I maintain that if MSC does in fact play a constitutive role in utterance interpretation then any explanatorily adequate semantic *theory* for natural language must account for the relevant aspects of a speaker's MSC. However, I readily grant that certain aspects of MSC, such as knowing how to spell a word, and which merely contributes to a speaker's functional literacy, is beyond the purview of semantic theory. Yet I will be focusing on the stronger thesis throughout—that at least one aspect of MSC—*viz*., our explicit knowledge of word-meanings—in at least some cases contributes directly to a speaker's "core" semantic competence and therefore constitutes a proper *explanandum* of semantic theory.

That's the project in a nutshell. The remainder of this introductory chapter is devoted to sharpening the relevant questions/issues and briefly sketching my proposed responses to them while postponing the finer details for later chapters. To help frame the proposal, I begin by trying to get clear about the fundamental goals of semantic theory, the nature of semantic competence, and the relationship between the two. Following the lead of a growing number of language theorists, I conclude that a semantic theory for natural language can more or less double as a theory of semantic competence. The relevant consequence is again that if MSC does in fact play a constitutive role in utterance interpretation, then any complete theory of semantic competence, and by parity of reasoning any explanatorily adequate semantic theory for natural language, must account for the relevant aspects of a speaker's MSC.

1.2. On semantic theory and semantic competence

Natural languages are distinguished by the expressions their native speakers utter and understand, where to understand an expression is pretty much by definition to grasp its meaning. Granted this characterization, the present challenge is twofold. First, one wants to know what this "grasping" consists in, which is traditionally the target of a theory of semantic competence. Second, one must specify the nature and structure of linguistic meaning; i.e., that which is presumably grasped when understanding occurs.[2] This latter question falls squarely within the domain of a semantic theory for natural language.[3] Or as more broadly construed Larson & Segal (1995: 1) suggest that "Ultimately, its goal is to provide theoretical descriptions and explanations of all of the phenomena of linguistic meaning." The wider aim of semantic theory, in other words, is to describe and ultimately explain the relevant *semantic facts* of understanding a language. So what are the relevant semantic facts? Well, this seemingly depends on the nature of linguistic meaning, which after millennia of debate is of course still far from being settled.[4]

As a way of escaping the impending circle, however, many language theorists nowadays seem willing to grant the following two more or less *a priori* claims:

(i) that linguistic expressions are the bearers of linguistic meanings, and

[2] What I am calling 'grasping' might itself be cashed out in any number ways; e.g., 'knowing', 'comprehending', 'cognizing', 'representing', or 'instantiating', to name but a few. But for the moment I will rely on our intuitive conception of what it is to understand the meaning of a linguistic expression.
[3] I will say more about the nature of linguistic meaning in Chapter 2.
[4] For example, most language theorists still operate under the traditional assumption that linguistic meanings are best characterized in terms of an expression's truth, reference, or satisfaction conditions. On this view, semantic theory is tasked, more specifically, with providing a recursive specification of the compositionally determined truth conditions for each sentence of the object language. Others, however, take meanings to be sets of Lewisian or Stalnakerian possible worlds, or alternatively structured Russellian propositions, or perhaps Fregean senses/thoughts. In each case the core *explanandum* is the relation between meaning and truth. In radical contrast, still others maintain that there are no such things as linguistic meanings, and thus nothing for a semantic theory to be about.

(ii) that linguistic meanings are the primary objects of understanding (i.e., semantic competence).

Of course, difficult questions remain regarding the exact nature of linguistic expressions themselves, and the ontological status of languages more generally. But under the plausible assumption that linguistic understanding can be characterized as a cognitive state of language users, it seems to follow that a semantic theory for natural language is again a theory of whatever competent speakers must know, or cognize, or otherwise mentally represent in order to grasp the meanings of their native language expressions.[5] If correct, it then becomes largely an exercise in empirical psychology and/or scientific epistemology to ascertain what, exactly, is the cognitive relation between speakers and the languages they are said to understand, and hence what, at bottom, semantic competence consists in. Furthermore, one assumes that the outcome of this endeavor will impose (possibly stringent) empirical constraints on the form that any explanatorily adequate semantic theory for natural language can take.

Here again difficult questions loom large. But according to an increasingly popular and empirically well-motivated movement among formal semanticists and philosophers of language, a semantic theory for natural language can perform double-duty as a theory of linguistic understanding, which is to say a theory of *semantic competence*. On its weakest construal, the task is to identify those semantic facts whose grasp would in principle suffice for an idealized speaker to understand the language under investigation. Others maintain, however, that a more explanatorily adequate semantic theory is one whose structure and content faithfully mirrors the structure and content of the semantic

[5] Although Wittgenstein evidently believed that understanding is *not* a cognitive process. Nor should being in a state of understanding be construed as a mental state. But frankly I don't know what sense can be made of such claims, as understanding is undoubtedly a property exclusive to cognitive creatures.

psychology of actual language users.[6] Yet whichever approach one finds most satisfying, somewhat deeper questions persist about the relation between speakers (whether actual or ideal) and the languages they are said to understand, or indeed whether semantic theory should even concern itself with such matters.[7]

Among those who *are* concerned with such matters, Michael Dummett notably argued that "To understand an expression is to *know* its meaning," as "an individual speaker's mastery of the language […] requires the notion of knowledge for its explication."[8] Upon first inspection Dummett appears to be using the locution "knowledge of language" in the traditional philosophical sense of possessing propositional knowledge of an expression's meaning, as traditionally cashed out in terms of knowledge of its truth, reference, or satisfaction conditions—what he sometimes refers to as a "full-blooded" theory of meaning.[9] Elsewhere, however, Dummett qualifies:[10]

> I shall reserve the phrase "a theory of meaning" for a theory thus conceived as something known by the speakers. Such knowledge *cannot be taken as explicit knowledge* [my emphasis], for two reasons. First, it is obvious that the speakers do not in general have explicit knowledge of a theory of meaning for their language; if they did, there would be no problem about how to construct such a theory. Secondly, even if we could attribute to a speaker an explicit knowledge of a theory of meaning for a language, we should not have completed the philosophical task of explaining in what his mastery of the language consisted by stating the theory of meaning and ascribing an explicit knowledge of it to him. Explicit knowledge is manifested by the ability to state the content of the knowledge.

---

[6] Advocates of this latter view are part of what Paul & Stainton (2009: 474) call the "New Philosophy of Language."

[7] Davidson (1986, 1990) held that the theoretically interesting relation holds between the *theory* (or theor*ist*) and the language under investigation as manifest in the interpretive *behavior* of its speakers—i.e., that a theory of meaning is a third-person description of the practice of using a language. However, I find such a view to be theoretically dissatisfying. And in spite of its potential historical/contrastive value, I will not consider it here.

[8] Dummett (1976:110, 113), my emphasis.

[9] Dummett (1974).

[10] Dummett (1976: 118).

As Dummett suggests here, ordinary speakers are typically incapable of articulating the compositional rules that govern assignments of meanings to expressions. Hence, these speakers presumably lack explicit propositional knowledge of any semantic theory that purports to describe and/or explain such knowledge. Rather, Dummett construes a speaker's mastery of a language as a *practical ability* "to employ the language on a certain occasion," yet which is nevertheless the manifestation of his/her *implicit* knowledge of meaning. For he adds (*ibid*., 118, 120):

> What an individual speaker's understanding of his language consists in is a legitimate philosophical enquiry; and it may be that, to explain this, we must invoke the notion of *implicit knowledge* […] This account can only be given in terms of the *practical ability* which the speaker displays in using sentences of the language; and, in general, the knowledge of which that practical ability is taken as a manifestation may be, and should be, regarded *as only implicit knowledge* [my emphases].

On the other hand, given his allegiance to Frege, Dummett is often accused of defending a thoroughly *anti-psychologistic* account of semantic competence—one which eschews appeal to the semantic psychology of actual language users. For there is a reading of Dummett according to which he is merely offering a rational reconstruction of language *use* as a social practice. In fact, I find Dummett difficult to pin down on this question (but then I am no Dummett scholar). However, as Antony & Davies (1997: 179) observe:

> Dummett often highlights what he takes to be the explanatory nature of a meaning-theory, encouraging the impression that he is looking for a psychologically realistic account of the means by which speakers master and deploy their languages.

Yet whatever is the correct interpretation of Dummett, so called "full-blooded" accounts of semantic competence face serious difficulties when it comes to explaining empirical facts related to language acquisition and thus how actual language learners ever manage to acquire and competently deploy the languages they are said to know/understand.

For this and other reasons, full-bloodedness has been most seriously challenged by Noam Chomsky's internalist/rationalist conception of linguistic competence as grounded in a tacit computational state of the mind-brain; *cf*. Chomsky (1986, 2000). More specifically, on Chomsky's view to "know" a language[11] is to possess an *I-language*—an *effective computational procedure* that generates infinitely many meaningful linguistic expressions from a finite stock of basic constituents according to a recursively specified set of grammatical rules.[12] In short, to know a language is to instantiate an I-language—a generative procedure, implementable by human biology, that enables its possessor to generate/represent and thereby understand the expressions of his/her native I-language. Linguistic understanding in turn depends on an interpreter's ability[13] to compute the meanings of expressions thus generated, which again constitutes what I am calling his/her linguistic-semantic competence (LSC).[14]

The knowledge/competence in question is characterized as "tacit" because again from a first-person perspective it manifests itself in a capacity to judge the various ways in which certain intrinsic semantic properties of expression-types constrain what their

[11] Notice that the locution "knowledge of language" carries no epistemic weight here as traditionally understood. But to avoid any unintended epistemological implications Chomsky nowadays employs the technical term 'cognize' as a replacement for the verb 'to know'. Thus, for readers who have qualms over my use of 'know' in this context feel free to substitute the technical term 'cognize'.

[12] Recursion (with respect to language) refers to the capacity to embed one clause, or complete sentence, within another in the same hierarchical (syntactic) structure, without limit, except as practically constrained by the limited cognitive resources of actual speakers. I will work through some of the details of Chomsky's notion of an I-language in the next chapter.

[13] *Pace* Dummett, by 'ability' I mean the exercise of a speaker's sub-doxastic linguistic competence, understood in the Chomskyan sense of biologically instantiating a stable computational state of the mind-brain, as opposed to a learned *practical* ability.

[14] While Chomsky does not distinguish linguistic competence from semantic competence (or what I am calling linguistic-semantic competence), this kind of distinction is often invoked by semanticists who believe that an independent compositional semantics can be given for natural language (i.e., a semantic theory divorced in certain respects from a theory of syntax). So distinguished, and as clarified below, I will assume that semantic competence is a proper subset of a speaker's core linguistic competence in Chomsky's sense. Thus, purists are free to omit the word 'semantic' in my use of the expression 'linguistic-semantic competence'.

token-utterances can (and cannot) mean. To borrow an example from Chomsky, consider the following sentences which on the surface differ only in the verbs ‘eager’ and ‘easy’:

(3) Max is eager to please

(4) Max is easy to please

Any competent speaker of English knows, on the basis of highly impoverished evidence, that (3) unambiguously means that Max is eager to please others whereas (4) unambiguously means that Max is easily pleased by others. Yet one can imagine a language in which (3) and (4) mean just the opposite, or are both ambiguous as between the two possible readings. For compare (5) which on the surface has the same basic form as (3) and (4) yet is two-ways ambiguous:

(5) Max is ready to please

Specifically, (5) has a reading according to which Max is prepared to please others and another by which Max is prepared for others to please him.

By common hypothesis these strings of words are assigned syntactic representations (or structural descriptions) that contain covert (unpronounced/unarticulated) syntactic elements called *traces*. Specifically, a standard analysis of (3) and (4) are given in (3') and (4'):

(3') Max_1 is eager [t_1 to please]

(4') Max_2 is easy [t_1 to please t_2]

In (3'), the idea is that the surface subject, ‘Max’, is initially generated as the grammatical subject of the verb ‘to please’. However, in order to satisfy certain structural constraints ‘Max’ is moved to the surface subject position leaving behind the trace ‘t_1’ that behaves like a silent pronoun and whose coindexation with ‘Max_1’ indicates their referential

codependency. The same goes for (4') except that in this case 'Max' is base-generated as the direct object of 'to please', indicated by 't_2', and where 't_1' represents an unpronounced grammatical formative that is tacitly understood to mean something like 'others', as in "others$_1$ to please Max$_2$."

This difference in syntactic configurations of (3) and (4) is again posited to explain their different yet *unambiguous* meanings. Similarly, the ambiguity of (5) is explained on the assumption that English grammar allows this particular string of words to be assigned more than one syntactic representation as posited in (5')/(5''),

(5') Max$_1$ is ready [t_1 to please]

(5'') Max$_2$ is ready [t_1 to please t_2]

where the same grammatical principles that govern the interpretation of (3') correspond with those of (5'), and where (4') likewise corresponds with (5''). Indeed, similar examples abound in all spoken languages. The upshot is this: While competent speakers seem to "know" such facts they are again generally incapable of reporting the linguistic principles that govern their interpretive judgments, *ex hypothesi* because these principles are "known" only tacitly (i.e., sub-personally/sub-doxastically).

In any case, I propose to not get bogged down here in a lengthy debate over the virtues of a Chomskyan conception of language, as nothing I have to say will win over detractors.[15] Rather, I will just adopt an I-language perspective without further argument. In regard to the question of semantic competence, I am disposed to follow the lead of theorists such as Evans (1981, 1982), Higginbotham (1989, 1991), Larson & Segal

[15] However, my core thesis, which I shall detail momentarily, is in many respects compatible with an externalist view of meaning. Furthermore, adopting an internalist/rationalist perspective of *language* is in principle no barrier to being an externalist about linguistic *meaning* (i.e., by individuating meanings externally). Indeed, there appears to be an ever-growing legion of truth-conditional philosophers of language who fit this mold. One recent example is Ludlow (2011).

(1995), Antony & Davies (1997), Smith (1992, 2006), and Pietroski (*forthcoming*), the consensus among whom is that appeal to the semantic psychology of actual speakers is indispensable in the construction of an empirically adequate theory of semantic competence. By parallel reasoning, Smith (1992: 114) adds that "What a plea for the relevance of psychological findings would ensure is that the form of a theory of meaning be treated as an empirical hypothesis." Taken a small step further, one is given to conclude that a semantic theory for natural language can, for most theoretical purposes, double as a theory of semantic competence. As such, one would expect a correct semantic theory to play a role in the explanation and hence theory of linguistic understanding.

1.3. On metalinguistic-semantic competence (MSC)

More to my immediate concerns, while one might agree with Chomsky that our knowledge of *grammar* is largely tacit, and thus generally unavailable to introspection, Smith (2006: 947-8) observes that "There is a vast amount of specific knowledge the speaker has and about which he is *authoritative*: including knowledge of what his words mean." And this is to suggest that "Our knowledge of word meaning is conscious and first-personal." Now, by authoritative Smith does not mean *infallible*. Rather, I take him to mean that, by and large, competent speakers have reliable and indeed quite often empirically justified beliefs about what their words mean. In particular, a speaker's explicit knowledge of word-meaning constitutes part of what I am calling his/her metalinguistic-semantic competence (MSC). Yet despite Smith's plea, MSC is typically dismissed by language theorists as a mere epiphenomenon—as being parasitic on the facts of knowing a language rather than partly *constitutive* of those facts. It is again the central aim of this study to demonstrate otherwise, and specifically against the backdrop

of a Chomskyan conception of words and their lexically-encoded meanings. But making this case requires first saying more about the nature of words and their lexical meanings.

1.3.1. A Chomskyan conception of words and their meanings

I will elaborate on these remarks in Chapter 2, but in brief outline Chomsky describes linguistic expressions as related collections of phonological, syntactic, and semantic properties, or "features" as they are often called. At the lexical level in particular we can think of word-meanings as *intrinsic* (again *non*-relational, language-*internal*) semantic properties of words that are tacitly understood by interpreters as linguistically-specified rules or "instructions" to activate one among possibly several semantically related *extralinguistic concepts*. The net result is that the meanings of words, so understood, guide and constrain without fully determining what those words can and cannot coherently be used to talk about in ordinary discourse.

Consider the word 'dog', for example, which typical English speakers initially lexicalize as an ordinary count noun and in turn *use* to express their concept of dogs; call this DOG(x). So understood, from the perspective of a speaker's semantic psychology, and to a rough first approximation, we might specify the meaning of 'dog' as follows,

(6) CON(√dog) = *activate*@DOG → DOG(x)

where '√dog' indicates the lexical root, as represented in a speaker's lexicon, corresponding to what we pre-theoretically think of as the word 'dog'. 'CON' is the name I give to a particular aspect of its context-invariant meaning, again as qualified in Chapter 2. The element '@DOG' in the metalanguage (i.e., on the right-hand side of the meaning assignment) represents a pointer to (address of) the concept DOG(x) in the speaker's long-term conceptual-semantic memory. In terms of comprehension, the element

activate@DOG is tacitly understood by interpreters as an instruction to activate their DOG(x) concept.[16] By hypothesis, however, DOG(x) is a language-*external* mental representation and thus strictly speaking not itself part of the meaning of the word 'dog' as encoded by its lexical entry. Rather, I include the expression '→ DOG(x)' as part of the specification of (6) for expository purposes to keep track of which concept typical English speakers associate with the meaning of 'dog'. And in the general case, the concept DOG(x) is located at the mental address specified by @DOG, where the arrow '→' should be read as "points to." In short, and strictly speaking, the meaning of 'dog' should be specified as (7),

(7) CON(√dog) = *activate*@DOG

where we can think of *activate*@DOG as ***the*** linguistically-encoded meaning of 'dog'. In turn, the meaning of 'dog' as specified in (7) will be understood as an instruction to activate the concept located at the address specified by @DOG. So qualified, I will nevertheless continue the practice of (6) for expositive clarity; the chief reason being that (*ex hypothesi*) for any given open-class word/lexical item there will typically be *multiple* concepts at the location specified by '@<*address-name*>'.

To briefly motivate this last point, observe that in appropriate contexts the word 'dog' can also be used as a transitive verb meaning something like 'to bother relentlessly', as in (8):

(8) Max dogged Minnie for a date until she finally relented[17]

[16] The notion of lexical meanings being understood as instructions to *activate* semantically related concepts is illustrative for present purposes. In chapter 5, however, I will trade in the notion of activation for one of *retrieving* or "*fetching*," to borrow the term of its inventor; *cf.* Pietroski (*forthcoming*).

[17] Ironically enough, as I was writing this chapter I noticed the following headline in the New York Times (02/17/2012): "Helen Radkey, researcher dogging Romney, has long questioned church's posthumous baptisms."

In this case, 'dog' (or the lexical root √dog) is used to express the two-place concept DOG(x, y), which is analytically related to DOG(x) but expresses a dyadic relation between an event and its two event participants. Moreover, I take it that most competent English speakers know that (8) is a legitimate use of 'dog', presumably based on its "core" noun-*ish* sense together with the linguistic context in which it occurs. In other words, I trust that most competent English speakers would readily endorse the following entailment:

(9) If Max dogged Minnie, then Max behaved as dogs sometimes do

This observation in turn suggests that the two senses of 'dog', while both formally and conceptually distinct, are analytically (or semantically) related, which to introduce a technical term is to say that they are *polysemes*.

In general, there is mounting empirical evidence to suggest that the semantic relationship between most if not all open-class words/lexical items and the concepts (or senses) they can be used to express in ordinary discourse is *one-to-many*.[18] In light of such facts, the meaning axiom posited by (6) above is more accurately construed as follows:[19]

(10) CON(√dog) = *activate*@DOG → {DOG(x), DOG(x, y)}

In this case @DOG serves as a pointer to a set or "family" of analytically related concepts in a speaker's long-term conceptual memory (again keeping in mind that DOG(x) and

[18] *Cf.*, e.g., Copestake & Briscoe (1995), Pustejovsky (1995), Klein & Murphy (2002), Beretta, et.al. (2005), Pylkkänen, et.al.(2006), and Klepousniotou, et.al. (2008). See also Chapter 5.

[19] I say *most* competent speakers because not every speaker lexicalizes the same concepts under the same words, where getting things "right" is part of what I above characterized as part of a speaker's conceptual semantic competence (CSC). However, as I argue in greater detail below, CSC in the sense of knowing what particular words *denote* is not directly relevant to a speaker's linguistic semantic competence (LSC). That is, one can deviate from local norm-governed standards regarding which words map to which concepts and still be a semantically competent language user, strictly speaking. The relevant claim, in other words, is that being semantically competent and being so regarded is not the same thing (more on this below).

DOG(x, y) are *not* strictly speaking part of the lexical entry for 'dog').[20] With respect to comprehension, (10) reflects the hypothesis that a typical utterance of 'dog' will be naturally understood by most competent English speakers as an instruction to activate one of *either* DOG(x) *or* DOG(x, y). What I am suggesting, in other words, is that while the word 'dog' has a single lexical entry with a univocal meaning it can nevertheless be used in different contexts to express different concepts (or *senses/polysemes*), though *as constrained by its linguistic meaning*.

One would of course like to know what determines which concept (or perhaps in some cases concepts) is selected for activation in a given context, which I take to be an open empirical question. However, the so-called process of "sense selection" doubtlessly involves both linguistic and extralinguistic cognitive mechanisms. More specifically, one assumes, in the general case, that the contextually-salient sense of an inherently polysemous word is determined in part by the linguistic context in which that word occurs and in part by the interpreter's recognition of the speaker's communicative intentions.[21] I will defend this hypothesis more thoroughly in Chapter 5. Yet for now I propose to set the question of sense selection to the side.

At this point I anticipate murmurs that I am losing touch with psychological reality. For if you ask an average English speaker what the word 'dog' means, he/she will surely not make reference to things like pointers, instructions, and concepts. Rather, a more intuitive response will be something to the effect that 'dog' *means*, or *denotes*, or *refers*

[20] Think here "family" in the sense of Wittgenstein's notion of "family resemblance," where in this context the relatedness of concepts is semantic, which is to say analytic.

[21] In the present example the linguistic context alone should suffice to disambiguate the intended sense of 'dog' according to whether it is used as a common noun or a transitive verb. However, as I qualify in the next chapter open-class lexical items typically carry additional semantic features that impose more fine-grained conditions on their possible interpretations.

to dogs (i.e., things that roll over and say "woof"). Indeed, I take it that such commonsense intuitions are in part what motivate semantic externalists to presume that lexical meanings can (and should) be specified in terms of a word's *truth*, *reference*, or *satisfaction* conditions, as indicated in (11):[22]

(11) $[\![\text{dog}]\!] = \lambda x.T$ iff x is a dog

The double-bar notation here indicates the semantic value of the enclosed expression as specified on the right-hand side of the assignment, which depending on one's theoretical commitments may or may not be identified with its *meaning*. In short, (11) can be read as stating that the word 'dog' is true of all and only those things in the domain of discourse that are dogs. However, if (11) is posited to represent or otherwise be strictly determined by the meaning of 'dog', then what are we to make of the equally commonsense intuition that the same word can also be used as a transitive verb meaning 'to bother relentlessly'?

A standard externalist rejoinder is to say that 'dog' is in fact *homophonous*, and more generally that the semantic relationship between words and their lexical meanings is strictly *one-to-one*. And this is to suggest that speakers encode two lexical entries with distinct meanings that have same phonology/pronunciation. To represent this hypothesis in externalist fashion we therefore need two lexical-semantic axioms, as in (12) and (13),

(12) $[\![\text{dog}_1]\!] = \lambda x.T$ iff x is a dog

(13) $[\![\text{dog}_2]\!] = \lambda x.\lambda y.T$ iff x dogged y

where the numerical subscripts reflect the presumed status of 'dog_1' and 'dog_2' as distinct words/lexical entries. This proposal is problematic for a number of reasons, however.

[22] Considerations of semantic compositionality are also typically cited as motivations for a truth-theoretic conception of meaning. However, semanticists such as Paul Pietroski (2005, *forthcoming*) have been making headway in demonstrating that truth-theoretic models are not the only way, and indeed perhaps not the most empirically adequate way, of constructing a compositional semantics for natural language.

First, if the entailment in (8) above is sound then 'dog' is clearly *not* homophonous as that notion is commonly understood but rather *polysemous*.[23] And as compared with (10), meaning axioms such as (12) and (13) fail to capture the semantic relatedness of these two senses of 'dog'. More generally, truth-theoretic meaning axioms fail to capture a range of relevant generalizations about the semantic relations between words that, by common measure, are proper *explananda* of semantic theory.[24] As described next, however, there is a way to accommodate externalist intuitions about a word's truth, reference, or satisfaction conditions yet without supposing that words are internally represented as encoding these conditions.

1.3.2. On the nature of metalinguistic-semantic competence (MSC)

What traditional meaning axioms like (12) and (13) come close to capturing—collectively, and save the biconditional—is a typical English speaker's considered *beliefs* about of the various *ways* in which the meaning of 'dog' constrains what it can and cannot be *used/uttered* to denote, refer to, or otherwise talk about in ordinary discourse. While the details here matter, the basic idea is straightforward. One of the few original yet less interesting claims of this work is that as speakers acquire new open-class words (i.e., lexical items), they also naturally acquire concepts *of* those words, or again to keep ideas straight what I will distinguish as *concepts*$_W$. Once acquired, speakers then use their concepts$_W$ to form explicit beliefs *about* the meanings of those words. More specifically still, and more interestingly, I suggest that explicit knowledge of or beliefs about lexical meanings are the conscious reflection of a speaker's tacit understanding of which

[23] Homophones are traditionally understood to have historically and/or analytically *unrelated* meanings.
[24] As I demonstrate in the next chapter, a Chomskyan-like specification of words and their lexical meanings is capable of capturing such facts.

concepts are coherently expressible with which words in which contexts, again *as constrained by their lexically-encoded meanings*. Collectively, these beliefs underwrite what I am calling a speaker's *metalinguistic-semantic competence* (MSC).

Consider, for example, that as a competent English speaker I know (or anyhow firmly believe) that in appropriate contexts the word 'dog' can be used, *inter alia*, either as a common count noun to designate dogs or as a transitive verb to mean something like "to bother relentlessly." In each case, I evidently used my concept$_W$ *of the word* 'dog' to form my explicit beliefs about what that word means, though specified in terms of what I take to be among its legitimate uses in ordinary discourse. Let's call this concept$_W$ $\underline{\text{DOG}}$(w) whose representational *content*, I propose, is the property of being the English word 'dog' as that word is instantiated in *my lexicon*. The relevant claim, in short, is that I have acquired the concept$_W$ $\underline{\text{DOG}}$(w) and used that concept$_W$, *inter alia*, to form certain explicit beliefs about what the word 'dog' means. But more specifically, I take it that my explicit beliefs about the meaning of 'dog' is the direct reflection of my tacit understanding of the various ways in which the linguistically-encoded meaning of 'dog', as encoded in my lexicon, guides and constrains without strictly determining which extralinguistic concepts that word can and cannot be used/uttered to express in ordinary discourse.

With respect to notation, underlining indicates that $\underline{\text{DOG}}$(w) is essentially metalinguistic in nature—it is again *ex hypothesi* a concept$_W$ *of* the word 'dog'. And the variable 'w' in $\underline{\text{DOG}}$(w) indicates, more specifically, that formally speaking it is a monadic predicate-concept that ranges over word-*tokens* as those tokens are represented

in a speaker's I-language/lexicon and in turn as physically manifest in verbal discourse, whether it be speech, sign, inscription, or what have you.[25]

Thus, to repeat, my considered beliefs about (or conception of) the meaning of 'dog' is a direct conscious reflection of my tacit understanding that the word 'dog' can in appropriate contexts be used/uttered to express *either* the monadic concept DOG(x) *or* the dyadic concept DOG(x, y), again as constrained by its lexically-encoded meaning. Such beliefs can be specified in a number of ways. But I find it illuminating to think of their mind-internal representations along the lines of (14) and (15) below, or what I like to think of as their *psychological forms*:[26]

(14) $\exists$w [$\underline{\text{DOG}}$(w) & EXPRESSES(w, DOG(x))]

(15) $\exists$w [$\underline{\text{DOG}}$(w) & EXPRESSES(w, DOG(x, y))]

By hypothesis, the higher-order concept EXPRESSES(*w*, *C*) designates a relation between a certain word, *w*, and a certain concept, *C*, that is tacitly known or otherwise believed to be coherently expressible with *w* in ordinary discourse. Thus, in paraphrase (14) reflects my tacitly held belief that the English word 'dog', which is to say the lexical item that my concept$_W$ $\underline{\text{DOG}}$(w) *represents* (or is a concept$_W$ *of*), can be used to express my concept DOG(x), whereas (15) reflects my belief that 'dog' can also be used in appropriate contexts to express DOG(x, y). In short, (14) and (15) together represent my tacit

[25] While one could construe concepts$_W$ as having singular contents that refer to word-*types*, I will assume for purposes here that concepts$_W$ range over word-tokens.

[26] This representation is admittedly highly speculative. I take it to be an empirical question what, precisely, are the underlying psychological forms of such beliefs. However, I trust that the formalism above conveys the content of my hypothesis clearly enough.

understanding that *either* DOG(x) or DOG(x, y), in appropriate contexts, can be expressed with *the same word*.[27]

Notice further that the word 'dog' can also be used *self-reflexively* in reference to the word 'dog' itself, as in (16),

(16) The English word 'dog' can be used to denote dogs

where to employ traditional terminology we can say that the word 'dog' is first *mentioned* and then *used*. Or equivalently, we can say that a speaker of (16) is using the word 'dog' twice over to express two different concepts.[28] Namely, the first occurrence of 'dog' expresses $\underline{\text{DOG}}(w)$ whereas the second expresses DOG(x). And from this it follows that both $\underline{\text{DOG}}(w)$ and DOG(x) are proper constituents of the belief expressed by (16), which is essentially that represented by (14). Put differently, what an utterance of (16) expresses is more or less equivalent to the belief represented by (14), which is again that DOG(x) is expressible with 'dog'. Similarly, (17) represents a subject's belief that 'dog' can be used self-reflexively to express $\underline{\text{DOG}}(w)$:

(17) $\exists w\,[\underline{\text{DOG}}(w)\ \&\ \text{EXPRESSES}(w, \underline{\text{DOG}}(w))]$

In other words, (17) corresponds to a subject's explicit belief that 'dog' can be used to refer to itself—i.e., to talk about the word 'dog'. In short, under present assumptions (14),

[27] I am not suggesting that such beliefs are *consciously* entertained by their subjects in terms of the semantic relations between concept-names and what these mental symbols express. Rather, from the subject's conscious perspective these beliefs are again understood in terms of what a given word can coherently be used/uttered to designate, refer to, or otherwise talk about in ordinary conversational situations.

[28] Compare sentences such as,

(i) Giorgione was so called because of size
(ii) Slim is so called because he is slim

where it appears that the names 'Giorgione' and 'Slim' are simultaneously used and mentioned.

(15), and (17) collectively constitute what we again might think of a typical English speaker's conception of what the word 'dog' means.[29]

The idea, quite generally, is that this sort of lexical polysemy is explicitly tracked by language users through the formation of multiple metalinguistic beliefs about which concepts are coherently expressible with which words in which contexts as constrained by their lexical meanings. As a way of capturing this most recent observation, I propose to extend the meaning axiom in (8) as follows:

(18) CON(√dog) = *activate*@DOG → {<u>DOG</u>(w), DOG(x), DOG(x, y)}

Specifically, (18) represents the fact that the context-invariant, lexically-encoded meaning of 'dog' facilitates the use of that word (in appropriate contexts) to express one of either <u>DOG</u>(w), DOG(x), or DOG(x, y). Another way of putting the point is that (18) is posited to indicate that its owner has *lexicalized* three concepts under (or with) the word 'dog', where the process of *concept lexicalization* establishes a fixed semantic relationship between a word's lexical entry and a veritable range of extralinguistic concepts that is known (or anyhow believed) to be coherently expressible with that word in ordinary discourse. In the reverse direction, speakers will naturally understand the meaning of a word as an instruction to activate one of the possibly many concepts that it lexicalizes, again as mediated and constrained by its lexically-encoded meaning.[30] In this way, I hope to have demonstrated how the meaning of a word can guide and constrain

[29] Of course, conceptions may vary from speaker to speaker and which may not always align with the norm-governed conventions of one's local linguistic community.

[30] I will summarize the details in Chapter 3, but I am assuming that all lexicalizable concepts, including metalinguistic concepts$_W$, are Fodorian atoms, which is to say syntactically unstructured, semantically primitive mental representations that express simple properties.

without strictly determining what that word can and cannot be used to denote, refer to, or otherwise talk about in ordinary discourse.[31]

Having introduced much of the technological apparatus that undergirds my core thesis, I devote the remainder of this chapter to a selection of potential applications of MSC while again postponing the finer details for later chapters. Specifically, I argue that MSC plays a constitutive role in the interpretation of syntactically simple proper names. Granted this assumption, I will then demonstrate how MSC can be used to explain Frege's classic puzzle about the informativeness of true natural language identity statements of 'α is β', where 'α' and 'β' are coreferential proper names. Lastly, I will attempt to show, more generally, how MSC is implicated in a speaker's *acquisition* of lexical meanings.[32] The upshot, again, is that should any of these proposals succeed then MSC can no longer be swept aside as a mere epiphenomenon but rather should be regarded a proper *explanandum* of semantic theory, and hence any complete/adequate theory of semantic competence.

1.4. On MSC and proper names

1.4.1. On the semantics of proper names

To my mind, the most obvious and direct role for MSC in utterance interpretation, and indeed in *expression* interpretation, arises in connection with proper names (or proper nouns, if you prefer) under the independently motivated assumption that names behave semantically as ordinary one-place predicates of individuals, which is to say predicates of *nameable things*, or of *things so-called*. Understood in this way, names are semantically

[31] As I argue in Chapter 4, a word-learner's conception of word-meanings will typically fluctuate over time as he/she sorts out which concept are coherently expressible with which words in which contexts.
[32] There are doubtlessly other interesting applications of MSC in the processes of utterance interpretation that I haven't the space here to explore and must therefore postpone for future study.

equivalent to common nouns (or cast in truth-conditional terms, expressions of type <*e*, *t*>). As I explain below, this is not to deny that names can be used/uttered to refer uniquely and rigidly to particular objects/individuals. Rather, I contend, in good company, that the otherwise institutionalized assumption that names are nothing more than labels for objects is disguised by the fact that when *used* referentially names combine, syntactically, with a covert index to form a complex *lexical expression*—and specifically an indexed noun phrase—that facilitates their use as devices of direct singular reference (i.e., as entity designators of type <e>).

While controversial, such a view is not without adherents; *cf*., e.g., Kneale (1962), Burge (1973), Katz (1994, 2001), Geurts (1997), Bach (1981, 1987 2002), and Larson & Segal (1995).[33] Specifically, I endorse a view according to which the meaning of a proper name (or proper noun) is best characterized as expressing the property of *being called by the name* '*PN'*, or perhaps more accurately being called by the *phonological form* of 'PN',[34] where we can think of "calling" as a kind of referring—i.e., as a kind of speech act. Now, if what I suggested earlier is on the right track, then like any other word the meaning of a name is understood by competent speakers as an instruction to activate a semantically related concept. And given present assumptions we might predict that the relevant concept is something like IS-CALLED(*n*) where the restricted variable '*n*' ranges over names. I think this is basically correct. However, instead of an individual variable

[33] For some cross-linguistic evidence to this basic conclusion, see also Izumi (2012).

[34] The qualification here is to suggest, in the first instance, that intuitively speaking "callings" are a kind of speech act done with sounds. In the second instance, being called by a name is not a property of its bearer, *per se*, but is rather a property of those who use that name in reference to its bearer (or bearers as the case may be). And while I use the expression "bearer of the name 'PN'," on my view while names have meanings they do not have bearers, or referents, as those terms are commonly understood. Rather, following others I take referring, like calling, to be a kind of speech act—something that people do with words/sounds, among other communicative devices such as pointings, winks, nods, and so forth.

my proposal is that the argument to IS-CALLED(_) is more usefully construed (or anyhow for certain theoretical purposes) as a concept$_W$ of the name in question.

By way of example, consider the name 'Aristotle' whose meaning, in my view, is best characterized as expressing the property of being called by the name 'Aristotle' (or perhaps the word-*sound* \ˈa-rə-stä-tˀl\).[35] Coupled with a Chomskyan conception of linguistic meaning, I propose to specify the meaning of 'Aristotle' as follows:

(19) CON(√aristotle) = *activate*@ARISTOTLE → IS-CALLED(ARISTOTLE(w))

As before, CON(√aristotle) is a way of representing a particular intrinsic semantic feature of the name/word 'Aristotle'. More specifically, by hypothesis the meaning of the name 'Aristotle' functions as a pointer, specified by '@ARISTOTLE', to the extralinguistic concept$_W$ IS-CALLED(ARISTOTLE(w)). With respect to comprehension, we can again think of '*activate'* as a lexical-semantic operation that instructs interpreters to activate one of possibly many concepts located at the address specified by '@ARISTOTLE', which given present assumptions includes the metalinguistic concept$_W$ IS-CALLED(ARISTOTLE(w).

To repeat, my hypothesis is that IS-CALLED(ARISTOTLE(w)) expresses the property of being called by the name 'Aristotle', or for brevity just *being called 'Aristotle'*. It then follows from its constituent structure that IS-CALLED(ARISTOTLE(w)) derives its content, in part, from the concept$_W$ ARISTOTLE(w), which again *ex hypothesi* expresses the property of *being the name (or word)* 'Aristotle'. I will again qualify these comments elsewhere. But in short my proposal is that is that (19) is way of representing what the name 'Aristotle' means. In turn, we can say that the concept$_W$ IS-CALLED(ARISTOTLE(w)) is what an *utterance* of the ***name*** 'Aristotle' expresses.

[35] I wish to remain neutral on whether it is the name itself or merely its phonological form that is most relevant to its linguistically-encoded meaning.

I emphasis ***name*** as such because like most other open-class words of natural language it strikes me as apparent that names have multiple *uses*, only one of which is to refer uniquely and rigidly to particular objects/individuals. Specifically, I assume that like 'dog' 'Aristotle' can also be used self-reflexively in reference to itself, as in (20):

(20) 'Aristotle' is used to refer to an ancient Greek philosopher

Notice, however, that in this context 'Aristotle' is not being used with its meaning, as such. Rather, it is being used to express its utterer's concept$_W$ ARISTOTLE(w), which is only *part* of the meaning of the name 'Aristotle'. In turn, I propose that the speaker's *belief* that ARISTOTLE(w) is expressible with 'Aristotle' can be specified as follows:

(21) ∃w [ARISTOTLE(w) & EXPRESSES(w, ARISTOTLE(w))]

In paraphrase, (21) represents a speaker's tacit belief that there exists a word, w, that is 'Aristotle', and that w can be used (in appropriate contexts) to express the concept$_W$ ARISTOTLE(w).

There are also contexts in which names can be used predicatively (or attributively) in reference to no one in particular, as with a sentence like (22):[36]

(22) Every Aristotle I know hails from Greece

(22) can be paraphrased as expressing the thought: For every *x* I know called 'Aristotle', *x* hails from Greece. Thus, I assume that what one expresses with 'Aristotle' in this context is simply its linguistically-encoded meaning as specified in (19), which again by hypothesis can be characterized as expressing the property of being called 'Aristotle'. Under present assumptions, in other words, I take it that the name 'Aristotle' as used in (22) expresses the metalinguistic concept$_W$ IS-CALLED(ARISTOTLE(w)).

[36] Strawson (1974:47) calls this use of a name "virtually self-quoting."

At this point I have suggested that ‘Aristotle’ can be used to express at least two distinct concepts$_W$, ARISTOTLE(w) and IS-CALLED(ARISTOTLE(w)), as exemplified by (20) and (22), respectively. And these two uses correspond, respectively, with metalinguistic and predicative uses of the name ‘Aristotle’. To reflect these two of ‘Aristotle’ uses we should update the meaning axiom in (19) as follows:

(23) CON(√aristotle) = *activate*@ARISTOTLE → {ARISTOTLE(w), IS-CALLED(ARISTOTLE(w))}

Thus, observe that if the present hypothesis is on the right track even proper names exhibit a degree of lexical polysemy. Or in terms of a speaker’s semantic psychology, I am suggesting that the meanings of names are associated with a range of extralinguistic concepts that those names can coherently be used to express in ordinary discourse. In general, for any syntactically simple proper name, ‘PN’, the range of expressible concepts will always include the metalinguistic concepts$_W$ PN(w) and IS-CALLED(PN(w)).

With respect to semantic competence, my proposal is that if (23) is at least roughly what competent speakers grasp when they understand the meaning of ‘Aristotle’, then from the perspective of a speaker’s semantic psychology, the ability to understand what ‘Aristotle’ means requires *at a minimum*:

(i) having recorded the name ‘Aristotle’ in one’s lexicon,

(ii) having generated the meaning axiom specified by (23), which entails

(iii) having a concept$_W$ of the name ‘Aristotle’, and

(iv) knowing, more generally, the semantic role that names play in the grammar, which includes knowing that the meaning of any syntactically simple proper name, ‘PN’, expresses the metalinguistic concept$_W$ IS-CALLED(PN(w)).

Stronger still, on my view (i)-(iv) is *sufficient* to understand 'Aristotle', and hence to be a competent user of that name. And this is to suggest that speakers can be competent users of a name such as 'Aristotle' without knowing *who* or *what* that name is customarily used to denote/refer to—that is, despite the fact that such speakers may not be *regarded* a competent user of that name.[37]

1.4.2. Referential uses of proper names

Now, this is again not to deny the commonsense view that names have referential uses. I readily agree, for instance, that 'Aristotle' can be (and commonly is) used in reference to a particular individual, such as the philosopher of Greek antiquity, the contemporary Greek shipping tycoon, and so on.[38] Or put in conceptual terms, I am happy to agree that 'Aristotle' can be used to express the singular concepts ARISTOTLE_{PHIL}, ARISTOTLE_{TYC}, and so forth. In these contexts, however, there is reason to believe that the name 'Aristotle' combines syntactically with a referential variable or index to form a *non-terminal lexical expression* that contains the name 'Aristotle' as a proper constituent.

Specifically, my proposal (which I accept largely on authority of notable experts)[39] is that when used referentially the name 'Aristotle' will occur as a constituent of a lexical expression (LE) such as (24),

(24) $[[_{\text{NP}}\ [_{\text{PN}}\ \text{Aristotle}]][x_i]]$

[37] This point plays a crucial role in my proposed solution to Frege's puzzle of informativeness as detailed in Chapter 7. Recall, however, that if Kripke is right a speaker can nevertheless succeed in using 'Aristotle' to refer to Aristotle so long as his/her referential intentions are in the right place.

[38] Precisely which 'Aristotle' is referred to in a given context is, I assume, determined by the speaker's referential intentions. I further assume that when the same name is used multiple times in the same context to refer to distinct individuals, tokens of the LEs that contain those names will be non-coindexed (i.e., they will combine with distinct Tarskian sequence variables).

[39] In particular, Baker (2003) and Fiengo & May (1998, 2006).

whose context-sensitive, compositionally-determined *denotation* (or *semantic value*) is specified below in terms of its Tarskian satisfaction conditions (where g is an interpretation function that assigns values to variables relative to a sequence, σ):

(25a) $g(x_i)^{\sigma} = \sigma(i)$, $i > 0$

(25b) $g([[_{NP}\ [_{PN}\ \text{Aristotle}]][x_i]])^{\sigma} = \lambda x$: x satisfies 'Aristotle' iff $x = \sigma(i)$ & $\sigma(i)$ is called 'Aristotle'

According to (25b), the semantic value of (24) relative to any sequence, σ, of entities in the domain is the i-th object in σ. The category label 'PN' attached directly to the name 'Aristotle' its tokens apply only to objects in the sequence/domain that are in fact so-called (without offering a theory of what it is *to be* so-called). In this way, when the name 'Aristotle' is interpreted as the proper constituent of an *indexed LE* it will be understood as referring uniquely and rigidly to a particular 'Aristotle'.[40]

However, under present assumptions—and this is crucial—in order to understand a referential/singular use of 'Aristotle' one must interpret the *bare* (i.e., *non-indexed*) nominal expression '$[_{NP}\ [_{PN}$ Aristotle]]' as having the meaning specified in (23) above. I will again be working through these details more carefully in Chapter 2. But the relevant claim is that when speakers use the name 'Aristotle' referentially they are at once expressing the property being called 'Aristotle'—i.e., the meaning of the *name* 'Aristotle'—and *using* that name to refer uniquely and rigidly to a particular contextually-salient individual. In turn, I contend that if listeners recognize the speaker's referential intentions as such, they will understand the name 'Aristotle' as having been used in just this way.

[40] When the same name is used multiple times in the same context to refer to distinct individuals, tokens of the LEs that contain those names will be non-coindexed (i.e., they will combine with distinct Tarskian sequence variables).

In summary, I see at least three distinct uses of a proper name. As just suggested, 'Aristotle' can be used referentially to rigidly designate a particular individual, in which case this use of 'Aristotle' will be understood, in part, as expressing a singular concept of the individual in question; e.g., ARISTOTLE$_{PHIL}$, ARISTOTLE$_{TYC}$, etc. In other words, we can think of singular uses of 'Aristotle' as expressing the property of being Aristotle-the-philosopher, or Aristotle-the-tycoon, and so on for each individual who is (or anyhow is believed to be) properly so-called. However, in the context of a sentence such as (20) 'Aristotle' can be also used to express the metalinguistic concept$_W$ ARISTOTLE(w), whose content is the property of being the name 'Aristotle'. Lastly, in purely predicative contexts such as (22), I take it that 'Aristotle' is used to express its context-invariant meaning, which by hypothesis speaker's tacitly associate with the concept$_W$ IS-CALLED(ARISTOTLE(w)) whose representational content is the property of being called [by the name] 'Aristotle'.

Furthermore, I have suggested that a speaker's ability to represent the meaning of 'Aristotle' requires having a metalinguistic concept$_W$ of the name 'Aristotle'. And while it may be true that speakers who have acquired the name 'Aristotle' (have recorded that name in their lexicon) are typically familiar with at least one its bearers, (23) indicates that being so-acquainted with one or more of its bearers is not a precondition of understanding the *meaning* of 'Aristotle'. This point bears emphasis because as Kent Bach (2002: 76) remarks:

> …one's knowledge about particular bearers of particular names does not count as strictly linguistic knowledge. Rather, it is in virtue of one's general knowledge about the category of proper names that one knows of any particular name that when used in a sentence (whether as a complete noun phrase or part of a larger noun phrase) it expresses the property of bearing that name.

Similarly, Segal (2001: 550) states that "It is part of our semantic competence to know, in general, what it is for something to bear a name." On my view, more specifically, to be a *linguistically competent* user of a proper name requires knowing only:

(i) what words of that grammatical category contribute to the meanings of the sentences in which they appear, and

(ii) how the meanings of those sentences restrict the range of thoughts they can be used to express.

And if what I said above is correct, what names contribute the meanings of sentences in which they appear is the property of being called by that name. If this were not the case it would be difficult to explain how we manage to acquire, deploy, and understand names (*a lá* Kripke, 1979) in the absence of a concept of who or what those names refer to. Indeed, I believe the principles in (i) and (ii) apply not just to proper names but to all open-class lexical categories.

To be clear, I am not denying that semantic theory is responsible for specifying the meanings of words, or what we might think of as their *I-meanings*.[41] Yet as understood here I-meanings are *purely formal* properties of expression-types that impose semantic constraints on how their token-utterances can (and cannot) be used in ordinary discourse. At the lexical level, there is good reason to suppose that many such properties are lexicalized when a word is acquired, and specifically those semantic properties that establish the initial boundary conditions on the range of concepts that corresponding concept-words can be used to express. These conditions in turn correspond to what appears to be a finite range of universal grammatical categories (or subcategories) that

[41] I am introducing the technical term 'I-meaning' here as a synonym for 'linguistic meaning' merely as way to briefly illustrate the nature of the latter. But hereafter I will dispense with the term 'I-meaning'.

determine each word's potential role within the language. With respect to names, these conditions again constrain the meaning of a name to things so-called.

I elaborate on these notions in subsequent chapters, but the present point warrants a brief digression. On the one hand, when it comes to theorizing about I-languages it is important to abstract away from individual or highly idiosyncratic differences that are largely irrelevant to the kinds of theoretically interesting generalizations that keep linguists awake at night. On the other hand, we must also guard against abstracting away from reality. For as James Higginbotham (1994: 165) aptly notes, "[semantic theory] cannot depend on an idealization of our capacities so extensive as to outstrip what is potentially available to us in thought."

As a general methodological principle, the empirical study of language as a natural object should however abstract away from contingent facts about:

(i) norm-governed linguistic conventions,

(ii) irrelevant variation between the open-class vocabularies of individual speakers, and in turn

(iii) pre-theoretical intuitions about what does and does not count as semantic competence.

Rather, again following Chomsky we should focus instead on universal linguistic principles—that which is common to all competent speakers across all naturally acquirable human languages. Indeed, evidence strongly suggests that at some theoretically interesting level of abstraction all languages rely on the same set of principles as opposed to "peripheral modifications."[42] And the most theoretically interesting questions turn on just those principles knowledge of which makes linguistic

[42] Chomsky (1986: 25).

understanding, and hence semantic competence, so much as even possible. As it concerns my project, the relevant questions relate to the knowledge/competence needed to acquire words with their linguistic meanings, and then deploy those words in grammatically acceptable ways as constrained by their meanings.

The basic moral is that we should not expect a semantic theory for natural language to specify the norm-governed extensions of individual words—this is the job of lexicographers. Granted, it may be nomologically impossible to acquire a language, and thus some core level of linguistic competence, without first acquiring a suitable array of its basic constituents. In particular, one must presumably acquire a representative selection of its open-class lexical categories (e.g., nouns, verbs, adjectives), along with its fixed logical/functional vocabulary (e.g., quantifiers, determiners, pronouns, prepositions, connectives/copulas, auxiliaries, and so forth). Yet once a language has been acquired—i.e., a generative grammar and a suitably primed lexicon—the depth and sophistication of one's open-class vocabulary is *not* a useful measure of one's linguistic competence, semantic or otherwise, or again at least not in any sense relevant to the kind of semantic theory under consideration here.

In short, what is minimally required of linguistic-semantic competence (LSC), specifically at the lexical level, is knowledge of a linguistically-specified rule or instruction for how to use words in accordance with constraints imposed by:

(i) their grammatical category (relative to the context in which they occur), and

(ii) their linguistically-encoded meanings, yet

(iii) only once their category information and meanings have been acquired, which is to say once that information has been assigned to words by their owner's I-language.

I will again say more about the nature of linguistic meanings in the next chapter. But returning from my digression, the upshot of the preceding section is that if MSC is required to understand the meanings of proper names, it follows that a theory of MSC is needed to augment any complete theory of a speaker's core semantic competence. Moreover, if one assumes, as I do, that a semantic theory for natural language can moonlight as a theory of semantic competence, it also follows that relevant aspects of MSC are legitimate *explananda* of semantic theory.

1.5. MSC and Frege's puzzle of informativeness

Utilizing resources introduced above, I attempt demonstrate in Chapter 7 how MSC might figure into a plausible solution to Frege's puzzle about the informativeness of true natural language identity statements of the form 'α is β', where 'α' and 'β' are coreferential proper names. The story here is also somewhat involved but let me attempt to briefly summarize.

As often described, Frege's puzzle of informativeness arises from the substitution of coreferential names in the context of identity statements. For instance, imagine a subject, Max, who has the names 'Hesperus' and 'Phosphorous' in his lexicon but does not know that these names corefer. However, Max does know *a priori* that (26) and (27) below are trivially true and thus thoroughly *uninformative*. Yet presumably Max can know this much without believing (28) *if* he fails to know (or anyhow believe) that there is just one object so-called.

(26) Hesperus is Hesperus

(27) Phosphorous is Phosphorous

(28) Hesperus is Phosphorous

Now, it obviously won't help to relieve Max of his confusion for a third-party informant to reassert the trivial truth of (26) or (27). By contrast an utterance of (28), if accepted as true, is putatively capable of rationally compelling Max to revise his mistaken belief about how many objects are the bearers of 'Hesperus' and 'Phosphorous'.

Now, in my view the only way to understand how an utterance of (28) can be informative for a subject in Max's epistemic condition is to first understand how he came to be in this condition in the first place. On this count, I think Fodor (2008) offers a highly plausible diagnosis of Max's dilemma as couched in purely psychological (i.e., conceptual) terms. To begin, let us assume that Max is capable of thinking about Hesperus *as such*, which is to say *as Hesperus*. Likewise, let's suppose that Max is capable of thinking about Phosphorous *as such*, which is to say *as Phosphorous*. On Fodor's view, this is to presume that Max is in possession of two formally distinct yet content-identical singular concepts; call them HESPERUS and PHOSPHOROUS. That is, by assumption both HESPERUS and PHOSPHOROUS *represent the same object*; namely, the planet Venus. There is, then, an intuitively clear sense in which Max knows *de re* which object is referred to with the names 'Hesperus' and 'Phosphorous'. Rather, the problem is that when using HESPERUS to think about Hesperus, and PHOSPHOROUS to think about Phosphorous, Max simply does not realize that he is in fact thinking about the same object. Max's conundrum, in other words, is that he simply fails to recognize that his concepts HESPERUS and PHOSPHOROUS are in fact content-identical. And *ex hypothesi* this because HESPERUS and PHOSPHOROUS, being formally distinct mental representations, play distinct computational roles in Max's mental economy. Moreover, it is presumably

for this reason that Max also does not know/believe that the names ‘Hesperus’ and ‘Phosphorous’ corefer.

From a semantic perspective, Max correctly believes that ‘Hesperus’ can be used (referentially) to express his concept HESPERUS, and that ‘Phosphorous’ can be used to express PHOSPHOROUS. And under present assumptions this is because Max associates the meaning of ‘Hesperus’ with HESPERUS and the meaning of ‘Phosphorous’ with PHOSPHOROUS. More specifically, and as indicated earlier, I assume that the meanings of the names ‘Hesperus’ and ‘Phosphorous’, respectively, as recorded in Max’s lexicon can be specified as follows:

(29) CON(√hesperus) = *activate*@HESPERUS → {HESPERUS(w), IS-CALLED(HESPERUS(w)), HESPERUS}

(30) CON(√phosphorous) = *activate*@PHOSPHOROUS → {PHOSPHOROUS(w), IS-CALLED(PHOSPHOROUS(w)), PHOSPHOROUS}

Notice that just like the rest of us, I assume that Max represents ‘Hesperus’ and ‘Phosphorous’ as having different meanings. In particular, and again strictly speaking, the meaning of ‘Hesperus’ is *activate*@HESPERUS and that of ‘Phosphorous’ is *activate*@PHOSPHOROUS.

I further assume that in order to explicitly track which objects/individuals are customarily referred to with which names, competent speakers use their concepts$_W$ of names to form metalinguistic beliefs about which concepts those names can coherently be used to express in ordinary discourse. For instance, by assumption Max knows/believes (again like the rest of us) that referential uses of ‘Hesperus’ express HESPERUS and that ‘Phosphorous’ expresses PHOSPHOROUS. And this is to suggest that Max has the following metalinguistic beliefs (again as tacitly/internally represented):

(31) ∃w [HESPERUS(w) & EXPRESSES(w, HESPERUS)]

(32) ∃w [PHOSPHOROUS(w) & EXPRESSES(w, PHOSPHOROUS)]

However, as a Frege-subject Max's dilemma is again that while his concepts HESPERUS and PHOSPHOROUS are in fact content-identical, he does not represent them *as being* content-identical. And this is due in part to the fact that Max does not know/believe that the names 'Hesperus' and 'Phosphorous' corefer. That is, part of Max's dilemma is that he falsely believes that the meanings of 'Hesperus' and 'Phosphorous' are *mutually exclusive* with respect to which object or objects they are customarily used to refer to.

Now, without taking a firm stance on precisely which thought/proposition is literally/conventionally *expressed* by an utterance of (28)—"Hesperus is Phosphorous"—the relevant question, in my view, is what misguided but otherwise rational and linguistically competent Frege-subjects such as Max stand to learn by *understanding* its utterance and in turn accepting it as true, say as based solely on the testimony of a knowledgeable and trustworthy informant. In Chapter 7 I argue that what Max learns from his acceptance of (28) is first that the names 'Hesperus' and 'Phosphorus' do in fact corefer. And from this proposition he can readily infer that just one object is a/the bearer of both names. Having accepted this much, Max will thereby be rationally compelled to tacitly conclude that his singular concepts HESPERUS and PHOSPHOROUS are in fact content-identical. However, as I argue in Chapter 7 the informativeness of the utterance rests in the derived proposition that one object is a/the bearer of 'Hesperus' and 'Phosphorus'. Importantly, I contend this much is available to *any* competent subject regardless of what he/she may or may not know about the relevant object of reference. Indeed, subjects can learn this much even when they are mistaken about or even fail to

know precisely who or what the names ‘Hesperus’ and ‘Phosphorous’ are customarily used to refer to.

For instance, suppose that Max not only positively disbelieves that the names ‘Hesperus’ and ‘Phosphorous’ corefer but mistakenly associates the meaning of ‘Hesperus’ with his concept of Mars, which is to say MARS. Let’s suppose further that Max misguidedly associates the meaning of ‘Phosphorous’ with his concept of Jupiter, or JUPITER. That is, I am supposing that Max harbors the following metalinguistic beliefs:

(33) ∃w [HESPERUS(w) & EXPRESSES(w, MARS)]

(34) ∃w [PHOSPHOROUS(w) & EXPRESSES(w, JUPITER)]

In this case, Max will be mistaken about which object (or objects) his informant is referring to with the names ‘Hesperus’ and ‘Phosphorous’. However, by accepting the truth of (28) Max will again come to believe, correctly, that the names ‘Hesperus’ and ‘Phosphorous’ corefer. And from the truth of this proposition he will also correctly infer that there is just one object under discussion that is a (or perhaps *the*) bearer of both names. Yet he will falsely infer, if only tacitly, that his concepts MARS and JUPITER are content-identical.

Nevertheless, as I argue in Chapter 7 an utterance of (28) can still be informative in the relevant sense for a Frege-subject in Max’s epistemic predicament.[43] For suppose that having accepted (28) as true Max goes on to learn, correctly, that ‘Hesperus’ is customarily used to refer to the planet Venus *qua* Hesperus, which is to say as the planet Venus appears in the evening sky. This is to suggest that Max comes to understand, correctly, that ‘Hesperus’ is used to express his concept HESPERUS, not MARS. In

[43] Actually, in Chapter 7 I use a slightly different example to the same conclusion. Specifically, I consider a subject who lacks the concepts HESPERUS and PHOSPHOROUS altogether.

consequence, however, Max will also come to believe *incorrectly* that 'Hesperus' and 'Phosphorous' are true synonyms—i.e., that both names express the concept HESPERUS as determined by their linguistically-encoded meanings. In other words, Max will come to associate the meanings of both 'Hesperus' and 'Phosphorous' with his concept HESPERUS. Yet notice that having learned that 'Hesperus' expresses HESPERUS Max will thereby know, *in virtue of his prior acceptance of* (28), that 'Hesperus' and 'Phosphorous' refer to the same object, which Max now correctly believes to be the planet Venus (*qua* Hesperus). Importantly, however, and to put the point counterfactually, under present assumptions Max would not have come to believe this latter proposition had he previously rejected the truth of (28). And given such observations I conclude that an utterance of (28) was informative for Max in the relevant sense despite the fact that he was mistaken about the identity of the object of reference.

To be in possession of HESPERUS and PHOSPHOROUS implies that Max is acquainted with the planet Venus, whether it be by direct ostension or indirectly by uniquely identifying definite description. There is, then, an intuitively clear sense in which Max knows, in a *de re* sense, *which* object is being referred to with the names 'Hesperus' and 'Phosphorous'. Rather, the problem is that when using HESPERUS to think about Hesperus as such, and PHOSPHOROUS to think about Phosphorous as such, Max does not realize that he is thinking about the same object; *viz*., the planet Venus. Max's conundrum, in other words, is that he simply fails to recognize that his concepts HESPERUS and PHOSPHOROUS are in fact content-identical. And it is presumably for this reason that Max also does not know/believe that the names 'Hesperus' and 'Phosphorous' corefer. Thus, what Max

presumably learns by his acceptance of (28) is that just one object is a/the bearer of both names.

Frege's puzzle then amounts to this: In virtue of what is (28) informative in a way that (26) and (27) are not? In general terms, Frege wondered why identity statements of the form 'a = b' in invented formal languages have a kind of "cognitive significance" that statements of the form 'a = a' and 'b = b' lack. Frege's solution to the puzzle is that proper names have both a *sense* and a *reference*, where the sense of a name, or in neo-Fregean terms its *meaning*, is an abstract mind-independent entity that uniquely determines its reference. So conceived, Frege reasoned that identity statements of the form 'a = b' express a relation between two different *senses* or "ways" in which type-distinct coreferential names determine (or present) their common referent. Thus, to grasp the informativeness of (28) is to grasp the fact that the respective senses of 'Hesperus' and 'Phosphorous' determine the same referent.

In purely psychological terms, it is widely assumed that for a Frege-subject such as Max to grasp the sense of 'Hesperus' he must be capable of representing Hesperus *as such*, which is to say *as Hesperus*. And to represent Hesperus as such seems to require having a concept of Hesperus *as such*, or what I will call HESPERUS. Likewise, to grasp the sense of 'Phosphorous' presupposes possession of PHOSPHOROUS. In general, it would seem that for Max to grasp the *informativeness* of (28), he must grasp the *thought* (or *proposition*) that it expresses (or rather its utterances relative to a context). And to grasp the thought expressed by (28) seemingly requires the capacity to mentally represent that thought as such. In turn, it would seem that to represent the thought expressed by (28) Max must possess the concepts HESPERUS and PHOSPHOROUS.

But more generally, the problem for Frege-subjects such as Max seems *not* to be that they fail to know *which* object is being referred to. Rather, their problem is that when faced with an utterance of (28) they simply fail to recognize that there is just *one* contextually-salient object that is a (or perhaps *the*) bearer of two names. Indeed, this could be true of subjects who are mistaken about the actual identity of Hesperus/Phosphorous yet who also mistakenly believes that the names 'Hesperus' and 'Phosphorous' do not corefer. Yet as I argue in Chapter 7, an utterance of (28), if accepted as true, can nevertheless be informative in the relevant sense for subjects in this epistemic condition.

Described in this way, my proposal in Chapter 7 is that what *any* competent speaker understands from utterances of true natural language identities of the form 'α is [the same as/identical to] β' is that there is there exists just one contextually-salient bearer of the names 'α' and 'β'. Importantly, this is the case regardless of whether or not Frege-subjects are acquainted with the relevant object of reference. For what any competent interpreter will know, based solely on its logical/linguistic form, is that for the utterance to be true it must be the case that the names 'α' and 'β' corefer. And from the presumed truth of this latter proposition interpreters can readily deduce that there is just one contextually-salient bearer of the names 'α' and 'β'. In this sense, my proposal is rather like a psychologized version of Frege's *Begriffsschrift* account of identity. This hypothesis again requires quite a bit of unpacking, but allow me another moment to develop the view employing current technology.

As indicated earlier, I am taking for granted that the meanings of proper names are essentially predicative (or metalinguistic) in nature. For example, I take it that the

context-invariant meanings of the names ‘Hesperus’ and ‘Phosphorous’ can be specified as follows,

(35) CON(√hesperus) = *activate*@HESPERUS → IS-CALLED(HESPERUS(w))

(36) CON(√phosphorous) = *activate*@PHOSPHOROUS → IS-CALLED(PHOSPHOROUS(w))

where by assumption HESPERUS(w) is a concept$_W$ of the (Anglicized) name ‘Hesperus’ and PHOSPHOROUS(w) is a concept$_W$ of the name ‘Phosphorous’. When used referentially, my hypothesis is that ‘Hesperus’ and ‘Phosphorous’ will occur as constituents of the lexical expressions (LEs) specified by (37) and (38) below:

(37) [[$_{NP}$ [$_{PN}$ Hesperus]][x_1]]

(38) [[$_{NP}$ [$_{PN}$ Phosphorous]][x_2]]

Notice that while the LEs in (31) and (32) are non-coindexed, nothing in the grammar either precludes or demands their coreference. Rather, facts about coreference between type-distinct (i.e., non-coindexed) LEs containing proper names are determined by extralinguistic facts about how those names/expressions are customarily deployed in ordinary discourse. Moreover, I take it that competent speakers are at least tacitly aware of these facts.

In turn, the compositionally determined meanings of the syntactically complex LEs in (37) and (38) above are specifiable in terms of Tarskian satisfaction conditions as follows:

(39) ‘[[$_{NP}$ [$_{PN}$ Hesperus][x_1]]]’$^{\sigma}$ is satisfied by $\sigma(x_1)$ iff $\sigma(x_1)$ is called ‘Hesperus’

(40) ‘[[$_{NP}$ [$_{PN}$ Phosphorous]][x_2]]’$^{\sigma}$ is satisfied by $\sigma(x_2)$ iff $\sigma(x_2)$ is called ‘Phosphorous’

Under present assumptions, the LE '[[$_{NP}$ [$_{PN}$ Hesperus]][x_1]]' will denote the planet Venus just in case Venus is the first object in the sequence, σ, and is called [by the name] 'Hesperus'. Likewise, the LE '[[$_{NP}$ [$_{PN}$ Phosphorous]][x_2]]' will denote Venus just in case the second object in the sequence also happens to be Venus, and is called 'Phosphorous'.[44] Recall that the category label 'PN' formally restricts the satisfaction conditions of nominal LEs to all and only those objects in the sequence that are in fact so-called (again setting aside the question of what it is *to be* so-called). In terms of comprehension, this restriction is enforced because in order to interpret the context-sensitive semantic values of the LEs in (33) and (34) one must again first interpret the context-invariant meanings of the names 'Hesperus' and 'Phosphorous'.

Now, in order to keep track of which objects/individuals are customarily referred to with which names, my hypothesis is that competent speakers use their concepts$_W$ of names to form beliefs about which singular concepts those names can coherently be used to express in ordinary discourse. For instance, recall Max who by assumption has the names 'Hesperus' and 'Phosphorous' in his mental lexicon, and whose meanings he represents as (35) and (36). Like the rest of us, Max knows that he can use 'Hesperus' referentially to express his singular concept HESPERUS, and that 'Phosphorous' expresses PHOSPHOROUS. And this is to suggest that Max has the following metalinguistic beliefs:

(41) ∃w [HESPERUS(w) & EXPRESSES(w, HESPERUS)]

(42) ∃w [PHOSPHOROUS(w) & EXPRESSES(w, PHOSPHOROUS)]

However, as a Frege-subject Max's dilemma is that while his concepts HESPERUS and PHOSPHOROUS are in fact content-identical, he does not represent them *as being* content-

[44] As a reminder, there is nothing in Tarski's system that precludes the same object from occurring multiple times in the same sequence.

identical. And this is due in part to the fact that Max does not know/believe that the names ‘Hesperus’ and ‘Phosphorous’ corefer.

Thus, we have generated a paradigm instance of Frege’s puzzle whose solution, given present assumptions, is fairly straightforward. First, imagine an informant—call her Minnie—who being aware of Max’s confusion utters (28), repeated below, and whose linguistic form is given in (28'):

(28) Hesperus is Phosphorous

(28') $[[_{NP}\ [_{PN}\ \textbf{Hesperus}]][\boldsymbol{x_1}]]\ [_{VP}\ [_{V}\ \textbf{is}]\ [[_{NP}\ [_{PN}\ \textbf{Phosphorous}]][\boldsymbol{x_2}]]$

Since there is just one contextually-salient bearer of the names ‘Hesperus’ and ‘Phosphorous’, the sequence variables $[x_1]$ and $[x_2]$ in the LEs ‘$[[_{NP}\ [_{PN}$ Hesperus]]$[x_1]]$’ and ‘$[[_{NP}\ [_{PN}$ Phosphorous]]$[x_2]]$’ will be saturated by the same object, *viz*., the planet Venus. And let us just assume for the sake of argument that (28) expresses the thought/proposition that Hesperus is numerically identical Phosphorous.

In terms of *comprehension*, given Max’s contrary belief about how many objects are the bearers of ‘Hesperus’ and ‘Phosphorous’, he will initially be puzzled by the content of Minnie’s utterance. As from Max’s perspective it appears to assert a contradiction. However, Max can readily infer that *whatever* is expressed by Minnie’s utterance, for her utterance to be true it must be the case that ‘Hesperus’ and ‘Phosphorous’ corefer. And from the truth of this propositions Max can tacitly infer that his concepts_W of the names ‘Hesperus’ and ‘Phosphorous’, which is to say HESPERUS(w) and PHOSPHOROUS(w), must be concepts_W of coreferential names. For notice that HESPERUS(w) and PHOSPHOROUS(w) will be fully activated and available in consequence of having interpreted the meanings of ‘Hesperus’ and ‘Phosphorous’. As importantly,

from this latter proposition Max can readily deduce that there is just *one* contextually-salient bearer of 'Hesperus' and 'Phosphorous', and not two as he previously believed. Finally, should Max find reason to accept Minnie's testimony as true, he will thereby be compelled to accept that his singular concepts HESPERUS and PHOSPHOROUS are in fact content-identical. Max will then be equally compelled to adjust his beliefs about how many objects are the bearers of the names 'Hesperus' and 'Phosphorous'.

My hypothesis, in short, is that an utterance of (28) is *informative* only to the extent that listeners such as Max recognize the metalinguistic proposition that it *pragmatically conveys*, which is that the names 'Hesperus' and 'Phosphorous' corefer. However, from his acceptance of this proposition Max can readily infer a *substantive* proposition regarding *how many* objects are presumed to be the contextually-salient bearers of 'Hesperus' and 'Phosphorous'. Indeed, I argue in Chapter 7 that the most natural way to interpret an utterance of (28) is that its utterer is making a statement *not* about the numerical identity of Hesperus and Phosphorous but rather a statement about *how many* objects are the bearers of the names 'Hesperus' and 'Phosphorous' (which are distinct propositions). The truth of this latter statement can again be deduced, more or less *a priori*, from the grammatical form (28). To this extent an utterance of (28) can be informative even for listeners who are ignorant of what the names 'Hesperus' and 'Phosphorous' are customarily used to refer to. For what such listeners will have learned is that there is just one contextually-salient bearer of both names. Listeners will also know henceforth that whatever is true (*de re*) Hesperus is also true of Phosphorous—

facts which can be quite valuable in trying to ascertain the precise identity of Hesperus and Phosphorous.[45]

More generally, my proposal about the informativeness (or "cognitive significance") of statements of the form 'α is β' is governed by the following principle:

> **(INF)ormativeness:** A true identity statement of the form 'α is β', where 'α' and 'β' are lexical expressions that contain coreferential uses of proper names, is *informative* just in case its utterance expresses or otherwise pragmatically conveys information that is sufficient to rationally compel a linguistically competent yet otherwise misinformed subject who understands its meaning, and believes it true, to revise his/her mistaken beliefs about *how many* objects are the bearers of those names. Parallel remarks apply to statements of the form 'α is *not* β'.[46]

On my view, the information pragmatically conveyed by an utterance of 'α is β' is the *metalinguistic fact* that 'α' and 'β' are coreferential proper names. A subject's grasp of this proposition will in turn license what I take to be the *substantive* (i.e., *material*) inference that there exists at most one contextually-salient object/individual that is the bearer of both names.

I will again defend this hypothesis more thoroughly in Chapter 7, but before concluding this Introduction let me briefly discuss one additional role for MSC in connection with the *acquisition* of lexical meanings, which I briefly defend in Chapter 4.

1.6. MSC and the acquisition of lexical meanings

I suggested earlier that knowing the norm-governed denotations of particular open-class words (in general) is not a component of a speaker's core semantic competence but that knowing how to acquire words with their meanings is. Moreover, as a Chomskyan I

[45] In effect, I argue in Chapter 7 for a psychologized version of Frege's *Begriffsschrift* conception of informative identity statements that does not depend on his sense/reference distinction, yet which allows that such statements express substantive (i.e., object-involving) thoughts/propositions.
[46] To be clear, (INF) is forwarded as an epistemological or pragmatic principle, not a semantic principle.

am committed to the idea that acquiring an I-language is part and parcel of one's core linguistic competence. Since the lexicon is considered an integral component of I-languages, broadly construed, it follows that the ability to acquire words with their meanings constitutes part of one's linguistic-semantic competence (LSC). I will again say more about this in Chapter 4, but as Carey (1978), Carey & Bartlett (1978), and other developmental psychologists report, word acquisition is often a gradual and thus graded process. For it seems that one can acquire a word in the sense of having lexicalized its phonological form, perhaps together with certain semantically significant grammatical features, yet without knowing precisely (or in some cases even vaguely) what that word denotes. Stated in purely psychological terms, the claim is that one can acquire a word without knowing precisely which *concept* (or *concepts*, as the case may be) the word in question can coherently be used/uttered to express in ordinary discourse.

So what then is a word-learner to do in the meantime? As often put in the acquisition literature, learners form *hypotheses qua beliefs* about what their words mean. For instance, given suitable background beliefs a competent learner who initially acquires the word 'tabulian' in the context of, say, (43) below might well form the belief expressed by (44):

(43) The forest fire in Palawan destroyed not only most of the native mahoganies but unfortunately most of the rare tabulians as well.

(44) The word 'tabulian' designates some kind of tree, or other.

That is, our subject will treat (43) as a working hypothesis about what the word 'tabulian' *means* in advance of having acquired a more definite concept of what tabulians *are*. The central claim of Chapter 4 is that generation of a belief along the lines of that expressed by (44) can facilitate one's acquisition of the meaning of words such as 'tabulian'.

Indeed, I go so far as to suggest that such metalinguistic beliefs often play a role in the acquisition of lexicalizable concepts such as TABULIAN(x)—i.e., concepts that we associate with the meanings of words and in turn use those words to express.

Notice that to represent (44) requires representing the *word* 'tabulian'. And under present assumptions to represent the word 'tabulian' presupposes having a concept$_W$ of that word—i.e., TABULIAN(w). Deferring details, my proposal, in short, is that a speaker's hypothesis about the meaning of 'tabulian', and which I liken to a Roschian *prototype* for the concept TABULIAN(x), will have roughly the following logical form:

(45) TABULIAN$_P$ = $\exists$x $\exists F$ [F(x) $\leftrightarrow$ TREE(x) & $\exists$w [TABULIAN(w) & SATISFIES(x, w) $\leftrightarrow$ F(x)]]

In loose paraphrase (45) can be read as expressing the word-learner's belief that there is some property F such that some entity x is an F iff x is a tree (of some as-of-yet unspecified kind), and x satisfies the word 'tabulian' iff $F(x)$. That is a lot to digest in one swallow, in part because it builds upon ideas developed in Chapter 3. But the relevant claim is that formation of a meaning hypothesis such as (45) requires the recruitment of a speaker's MSC. The relevant consequence is this: If the acquisition of lexical meanings requires the exercise of one's MSC, then once again a theory of MSC is needed to complement a complete theory of LSC, and by parallel reasoning any explanatorily adequate semantic theory.

1.7. Closing remarks and next steps

The claim that speakers have explicit concepts$_W$ of and beliefs about the meanings of words is again nothing new or terribly profound. However, I am aware of only a few serious attempts to explore in any detail how MSC might contribute to utterance

interpretation and thereby constitute part of a speaker's core semantic competence.[47] Yet given the paucity of empirical evidence my conclusions are in some respects highly speculative. However, I believe that in all cases they are intuitively plausible and, moreover, fully compatible with the best available empirical data. As importantly, I think the proposal has wide-ranging explanatory power, with application to other philosophical issues that I haven't the space here to address.

At bottom, it strikes me as indisputable that MSC provides an important epistemic bridge between language and thought. Broadly speaking, our concepts$_W$ of words and the MSC they subserve provide a window, albeit often dimly lit, onto the intrinsic semantic properties and relations of lexical items as manifest by their physical tokens in ordinary discourse. So too, this study can be viewed as a somewhat dimly lit journey into the semantic psychology of utterance interpretation as viewed through the window of our explicit metalinguistic knowledge of/beliefs about the meanings of words. Its rather modest goal is to simply work through some of the finer details that might render my proposed solutions to the questions posed at least moderately compelling. More optimistically, my long-term goal is to generate enough cross-disciplinary interest in the project to prompt further empirical research that might help to either confirm or disconfirm the philosophically interesting aspects of my core thesis.

In outline of work to follow, I shall again be arguing that not only does a speaker's MSC often play an important role in utterance interpretation, but as just discussed that it has a role to play in the acquisition of lexical meanings. This latter topic is again the

[47] I have in mind here (i) Murphy (2003), specifically with respect to lexical-semantic relations such as antonomy, heteronymy, etc., and to whom credit goes for supplying the original inspiration for this project, and (ii) Fiengo and May (2006) with respect to the semantics of proper names. Chapter 6 is devoted to review (and criticism) of the latter.

focus of Chapter 4, and the former is the subject of Chapters 5, 6, and 7. Specifically, Chapter 5 addresses the role of MSC in the representation and contextual disambiguation of lexical polysemy. And to see just how far my thesis can be pushed, Chapters 6 and 7 explore how MSC might possibly stand to explain the informativeness of natural language identity statements involving coreferential proper names, and more generally failures of substitution of such names, *salva veritate*, in so-called "opaque" contexts.[48] Importantly, if even just one of these claims pan out, this result would stand to support my general conclusion that a theory of MSC is needed to supplement any explanatorily adequate semantic theory for natural language—a theory whose central goal, I again assume, is to specify what one must minimally know, or cognize, or otherwise mentally represent in order to be a competent speaker of one's native (I-) language.[49]

The immediate next steps (Chapters 2 and 3) are to lay bare several core background assumptions about the nature of human language, words, lexical meanings, lexical concepts (in general), more specifically what I am calling concepts$_W$ and the beliefs they are used to form. This initial exercise is designed in part to (i) level-set on some technical jargon, (ii) clarify relevant background assumptions, and thereby (iii) pave the way toward a proper understanding of my core thesis. Indeed, what may strike some readers as boring review in Chapters 2 and 3 is actually doing much of the heavy lifting in work that follows. This initial detour requires setting much of what I have just said on the back burner for later chapters. Yet I worry that anything of philosophical

[48] In retrospect, I am now of the opinion that no student of philosophy should be allowed to publicly comment on Frege's puzzle until they have been in the business for at least ten years!

[49] Among applications not covered here is the potential role of metalinguistic knowledge/competence in the interpretation of quotation expressions; *cf*., e.g., Cappelen & Lepore (2007). In addition, MSK/C doubtlessly plays a role in general literacy and bilingualism, which has been explored to some extent by others; *cf*., e.g., Sharwood-Smith (2004, 2010), Bialystok (2001), to name just two.

interest will only make sense as situated within an operational model of the relationship between language and thought. And while many such presuppositions are not, strictly speaking, critical to the success of my core thesis, I think they nonetheless sound given what is presently known about the nature of language and thought. I will formally conclude this study in Chapter 8 by reiterating the explanatory role of MSC.

2. On the Nature of Words and Their Lexical Meanings

2.1. Introduction

As this investigation has much to do with words, and in particular their lexically-encoded meanings, let me begin in earnest by saying something about what I take a "word" of natural language to be. And this is to specify, in part, the formal properties of their lexical entries that determine the identity of their tokens.[1] Doing so is no easy task, however, as the criteria must apply not only to words in current circulation but to all possible words across all naturally acquirable human languages. It is also somewhat paradoxical. For on the one hand competent speakers have an intuitively clear grasp of what words are. Or at least in clear cases we are generally adept at distinguishing words of our native language from, say, non-words, nonce words, foreign words, and multi-word phrases/idioms, as evidenced by our shared judgments on ordinary lexical decision tasks. On the other hand, possible exceptions and putative counterexamples abound, cross-linguistically, making it quite difficult to pinpoint a univocal definition of the word 'word'.[2] In these more troublesome cases, not only do ordinary intuitions quickly break down but even the experts—i.e., those in the "word" business—quite often disagree.

Despite these difficulties, most linguists and language-oriented cognitive psychologists typically take the existence of words for granted. However, widespread disagreement and/or lack of convincing arguments and/or hard evidence has fostered skepticism among some language researchers, and philosophers in particular, about the

[1] As Aydede (2000) rightly comments: "if you are advancing a theory that quantifies over certain entities like symbols realized in the brain, you owe an account of their individuation conditions.

[2] Indeed, according to some reports not all languages even have a word that means what the English word 'word' means.

psychological reality of words as a stable class of linguistic entities.[3] While the skepticism is in some sense justified, there is also reason for optimism. As to my mind this is neither exclusively nor centrally a question of how we ordinarily (or even philosophically) conceive of words. Rather, it is a substantive empirical question about how the human language system, understood as an object of natural inquiry, individuates the expressions that *it generates*. This is simply a reminder that with respect to the scientific study of natural language, and in the absence of compelling counterarguments, words are justifiably treated as *linguistic kinds* not unlike other scientific kinds—albeit kinds whose underlying nature has yet to be fully unearthed.[4] Progress has and currently is being made, however, particularly in the fields of language development/acquisition and psycholinguistics. The bottom line is that enough is now known to at least tightly constrain the range of plausible hypotheses.

So qualified, the aim of this chapter is to offer not a highly refined *theory* of words but rather merely a working characterization of what words must be like, or are most likely to be, in order to explain our intuitive judgments about them along with a representative range of well-attested linguistic facts. Doing full justice to this topic warrants a dissertation in itself. So to keep the discussion manageable, what follows amounts to a relatively brief and highly selective report of what I consider to be the most promising proposal, again given what is presently known. As readers may have guessed, I will be endorsing a broadly Chomskyan conception of words, though as qualified below.

---

[3] *Cf.* Lepore & Hawthorne *On Words* (2011) and Kaplan (1990) for useful discussion. By contrast, while linguists recognize the difficulty in specifying the identity conditions of words, they normally take their existence for granted.

[4] As such, it should come as no surprise that ordinary speakers, let alone trained theoretical linguists, often lack clear intuitions about the underlying nature of words. Indeed, it is a separate question how language users *conceive* of the intrinsic properties and relations of the words they utter. I broach this latter topic near the end of Chapter 3 and again in Chapter 4.

Once complete, this characterization will then serve as a *tentative* background assumption in all that follows. I stress *tentative* because I trust that my core thesis is largely compatible with other naturalistic outcomes in the same general vicinity. Settling on just one, as I will here, is largely aimed at facilitating subsequent discussion. While I also trust that most readers are at least vaguely familiar with Chomsky's wider project, allow me to first briefly elaborate on that enterprise, if only to introduce some terminology.

2.2. A Chomskyan framework for natural language

It is by now (or at least to my mind) an indisputable fact of human nature that all neurotypical children are born with an innate predisposition to acquire, under the pressure of limited and imperfect experience,[5] one of many possible human languages. If Chomsky is right in the details, this potential emerges from a biologically determined set of principles—a *Universal Grammar* (UG)—that governs the development (or *growth*, to use Chomsky's term)[6] of all naturally acquirable human languages. Where variation exists, language-specific constraints are thought to become activated/programmed in response to linguistic stimuli absorbed from the learner's ambient environment. Yet whichever variant is the target of acquisition, the end result is characterized as a stable, computational state of the child's mind-brain that Chomsky again calls an *I-language*, where the "I," he adds, "is chosen to suggest that the system is internal, individual, and

[5] As Chomsky (1965: 4) observes, children acquire languages as a result of being exposed to the speech performances of other speakers, which are characterized by "numerous false starts, deviations from rules, changes of plan in mid-course, and so on." Chomsky adds that "The problem for the linguist, as well as for the child learning the language, is to determine from the data of performance the underlying system of rules that has been mastered by the speaker-hearer and that he puts to use in actual performance."

[6] This is probably old news, but Chomsky quite literally thinks of human language as a biological organ (albeit mental in nature) which is "grown" in much the same way that chordates grow hearts and renates grow kidneys.

intensional."[7] I will say more about this below, but in short a theory of grammar is for Chomsky a theory of I-languages and how they are acquired.[8]

Just briefly, I-languages are to be contrasted with what Chomsky calls public *E*-languages where the 'E' here stands for "external" or "extensional." E-languages are what we normally associate with the world's spoken languages such as Russian, Japanese, Hindi, Urdu, Mohawk, and so forth, and their corresponding norm-governed grammars as often taught in elementary schools. From a theoretical perspective, E-languages are typically modeled as functions in *extension*—i.e., sets of ordered pairs that map publicly tradable E-expressions onto "entities drawn from some stipulated domain."[9] Understood in this way, however, E-languages are *not* by Chomsky lights proper subjects of natural scientific enquiry. An E-language he says:

> …has no particular status in the empirical study of language; it is something like phenomenal judgment, mediated by schooling, traditional authorities and conventions, cultural artifacts, and so on.[10]

In other words, E-languages are rooted not in human biology but in the *corpora* of expressions that people actually (or might possibly) utter as determined by the various and sundry (unpredictable/unscientific) ways that humans put language to use in pursuit of their life interests. However, such pursuits have little to do with *language*, *per se*, and a lot to do with the normative aspects of behavioral and social psychology. Rather, as

---

[7] Chomsky (2000a:120). That is to say, in one sense I-languages are largely *private* to their owners. However, *theories* of I-languages are intended to describe the psychology of an idealized speaker-hearer thereby abstracting away from the irrelevant vagaries of individual speakers.

[8] As mentioned in Chapter 1, once acquired the possession of an I-language (a natural language grammar), on Chomsky's view, constitutes the conditions for "knowing" or "cognizing" a language, which manifests itself in a speaker-hearer's linguistic competence.

[9] Chomsky (2000a: 39). More specifically, externalists usually specify sentential meanings in terms of a sentence's truth-conditions, which themselves correspond to propositions, or sets of possible worlds, or what have you.

[10] Chomsky (2000a:28).

Chomsky sees it the theoretical interest lies in the "completely internalist" aspects of human language, which is to say *I*-languages—mental "organs" that biologically implement functions in *intension*—where "questions of truth and falsity arise for grammar as they do for any scientific theory."[11]

I will return to the relevant details below, but in brief overview a Chomskyan I-language is characterized as a domain-specific module of the mind that consists in a *computational procedure* and a *lexicon*.[12] The lexicon is essentially a repository for individual "words," or "word-like units," of a given speaker's idiolect. Informally, Chomsky describes words as "arbitrary associations of sound and meaning."[13] In technical terms, each *lexical item*, or *LI* for short, is a complex mental representation that consists in related "bundles" of phonological and semantic "features" to which Chomsky assigns the labels PHON and SEM, respectively.[14] The notion of a "feature," Chomsky tells us, "is just the technical term for the elementary constituents of LIs and expressions constructed from them." Thus, on Chomsky's view it is lexical *features* and not *words*, *per se*, which serve as the basic primitives of natural language.

I will also say more about the nature of lexical features below, but to a first approximation we can think of Chomskyan words as <PHON, SEM> pairs—i.e., related bundles of formal lexical features that are interpretable by language-external cognitive

[11] Chomsky (1986: 22). Indeed, *qua* biological organ Chomsky often notes that I-languages are fundamentally systems of *thought* and only contingently useful for the externalization of thought in the service of interpersonal communication.
[12] Chomsky (1995: 33).
[13] To say that the mappings between word-sounds and their meanings are *arbitrary* is to just say that these mappings do not follow from general linguistic principles. Rather the meanings of individual words must be individually learned/memorized. As such, Chomsky often describes the lexicon as "a list of irregularities" or "idiosyncrasies."
[14] Chomsky (2000a: 125) writes "We may take the semantic features S of an expression E to be its *meaning* and the phonetic features P to be its *sound*; E *means* S in something like the sense of the corresponding English word, and E *sounds* P in a similar sense, S and P providing the relevant information for the performance systems." [emphases are Chomsky's].

systems. However, Chomsky also routinely adverts to certain purely language-internal features of words that are specifically relevant to lexical morphology and/or phrasal syntax yet which are not themselves interpretable at the interfaces to language-external systems.[15] Following others, I will group these so-called "uninterpretable" features of words under the label 'SYN'.[16] Thus, to a closer approximation we can think of Chomskyan words/LIs as particular instantiations of the complex mental structure <PHON, SYN, SEM>, which is to say related bundles or sets of phonological, morphosyntactic, and semantic features. A Chomskyan lexicon can therefore be characterized as the totality of such representations in a given speaker's idiolect.

The computational procedure, by contrast, or what is more commonly known as simply *the grammar*, is effectively a parametric instantiation of UG—a *generative procedure* instantiated in the mind-brain whose job it is to select items from the lexicon and combine them, recursively, according to the rules of grammar to *generate* increasingly more complex linguistic expressions (i.e., phrases and sentences). Importantly, expressions thus generated must satisfy conditions on interpretability (or "legibility") imposed at the interfaces to language-external performance systems, or what Chomsky calls the *articulatory-perceptual* (A-P) and *conceptual-intentional* systems (C-I), respectively.[17] When these output conditions are met the derivation is said to "converge" at the interfaces. Otherwise it "crashes" and the derived expression is judged

[15] With respect to speech production, it is merely a contingent fact about the human language system that it happens to be useful for verbal communication by making its outputs available to articulatory systems.
[16] While Chomsky does not himself make explicit reference to 'SYN' features, I doubt that he would object to those who do. For example, Chomsky (1965: 214, fn.15) writes "We might, then, take a lexical entry to be simply a set of features, some syntactic, some phonological, some semantic." And in his discussion of Grimshaw's (1981) notion of "canonical realization structures," Chomsky (1995: 32) writes: "Notice that this consideration indicates that lexical entries contain *at some syntactic information*…" [my emphasis].
[17] As Chomsky (2000a: 9) puts it, these are "systems that make use of the resources of the faculty of language…"

by competent speakers to either be ungrammatical, uninterpretable, or otherwise unacceptable. However, notice that while some word strings may not correspond to generable linguistic expressions they may, with extra cognitive effort, still be interpretable by other means; e.g., by means of reasoning, perhaps of the pragmatic sort, about the utterer's communicative intentions in saying what he/she did.

Before further evaluating Chomsky's notion of words *qua* lexical items, it will help facilitate discussion here and elsewhere to briefly mention a few other general points about his conception of I-languages. Observe first that the computational procedure is characterized by Chomsky as a monolithic combinatorial system. However, other linguists working in the generative tradition typically assume a division of computational labor between *morphology*, *phonology*, *syntax*, and *semantics*, treating each system as an independent sub-module of the grammar with its own well-defined interface (some of which are wholly language-internal). In brief, the morphological subsystem is said to be responsible for word formation—i.e., the assembly of lexical features into lexical items/LIs, and of LIs into morphologically complex word-forms (e.g., their inflected and/or derived forms). The syntactic system is then responsible for combining items so-assembled into more complex syntactic structures—i.e., linguistic expressions whose lexical constituents bear structural relations to one another. The job of the semantics module, by contrast, is to map these expressions onto their phrasal/sentential *meanings*, or *LFs* as they are called, which again must be *interpretable* at the interface to C-I. Operating in parallel, the phonological subsystem generates complex *phonological forms* (*PFs*) that must be *pronounceable* by the articulators via the A-P interface.

As I understand, the chief motivation for distinguishing morphology from syntax is that there is some reason to think that morphologically complex words are generated within the lexicon itself as opposed to the wider syntax, corresponding to what Chomsky has dubbed the *Lexicalist Hypothesis*, or what is also known as *Lexicalism*.[18] While opinions here vary, more recently Chomsky (1995: 20) himself proposes that only derived word-forms (such as the deverbalized noun 'destruction') are generated by processes "internal to the lexicon," whereas inflected forms (e.g., the regular past tense verb 'destroyed') are the products of "computational operations of a broader syntactic scope." But either way, it is widely agreed that once generated many morphologically complex word-forms can be manipulated (e.g., moved, merged) by phrasal syntax as atomic/indivisible units.[19] By contrast, the motivation for distinguishing phonology from syntax and semantics stems from the hypothesis that the linear surface structures of complex expressions (i.e., their pronounced phonetic forms) are quite different from their underlying logico-syntactic forms; their respective constituents not only differ in kind but stand in distinct structural relations to one another. And this is again because *ex hypothesi* the same expressions must satisfy two distinct interface requirements—they must again be both phonetically pronounceable and semantically interpretable. Different interface requirements in turn imply a grammar that generates two distinct structural representations (or *structural descriptions* as Chomsky calls them)—a *PF* and an *LF*—

[18] Chomsky (1995: 35) writes: "The primes constituting the terminal string of a phrase marker are drawn from the lexicon; others are *projected* from these *heads* by operations of the computational system." [Chomsky's emphases]. The *Lexicalist Hypothesis* is widely attributed to Chomsky (1970). For a recent defense, see Williams (2007). However, LH has been recently challenged by a movement known as *Constructivism*, or *Constructionism*, which I'll say more about below.

[19] Chomsky (2005a: 3) writes, for example: "Adopting the P&P framework, I will assume that one element of parameter-setting is assembly of features into lexical items (LIs), which *we can take to be atoms for further computation*…" [my emphasis]. This remark will however be qualified below.

the former consisting in a complex array of interpretable phonological features (PHONs) and the latter an array of interpretable syntactico-grammatical features (SEMs).

While it may be convenient for certain theoretical purposes to speak of morphology, phonology, syntax, and semantics as independent processing subsystems of the wider grammar, these may be just different ways of describing the same combinatorial system. Indeed, Chomsky nowadays recognizes just two basic grammatical processes (or operations): "one assembles features into lexical items, the second forms larger syntactic objects out of those already constructed, beginning with lexical items." In the wake of Chomsky's Minimalist Program (Chomsky [1993, 1995]), the derivational process is said to diverge at certain stages called "phases" where interpretable fragments of the emerging structure are "spelled out" by first "checking" (deleting) uninterpretable features (or "valuing" unvalued features) to ensure what Chomsky calls *Full Interpretation* (FI) at the interfaces (Chomsky [1984: 98]). One fragment, a PF, clothed only in phonological material, is forwarded to A-P for purposes of speech production (in the downstream direction). The other fragment, an LF, clothed only in logico-syntactic material, is sent to C-I (in the upstream direction) to be semantically interpreted. The bottom line is that the output of the computational procedure, whatever it consists in, must satisfy these two basic interface requirements.

With these preliminaries in place we can begin to more carefully formulate an answer to the question of what a "word" of natural language is, or rather again what they are most likely to be given our present state of knowledge. As a reminder, the relevance of this inquiry with respect to my core thesis is that a proper understanding of word-concepts, or again $concepts_W$ to keep ideas straight, requires an understanding of what

these concepts$_W$ are posited to be concepts$_W$ *of*, which I assume are *words* of natural language. More specifically, my primary interest here is with the nature of lexical meanings, or SEMs to use Chomsky's terminology, that speaker's use concepts$_W$ to form explicit beliefs about, though again as evidenced by the constraints they impose on what their host words can and cannot be coherently used to talk about in ordinary discourse. And from a theoretical standpoint one can only understand the nature of lexical meanings against the backdrop of a plausible general analysis of what words themselves are.

While certain aspects of my core thesis again do not necessarily rely on one particular analysis, the present exercise will provide a vocabulary with which to entertain questions that are directly germane to my core thesis. The issues here are subtle, and so it will pay to proceed slowly. In particular, I dwell on what has become known in linguistics circles as "feature theory," which is central to Chomsky's conception of words, and of linguistic expressions more generally, as complex non-conceptual mental structures/representations. This discussion will help in particular to clarify the nature of the lexical meaning axioms postulated in Chapter 1, which elaborates on Chomsky's notion of SEM features to include what there I called 'CON'. In addition, the notion of a "feature" will reappear in the next chapter in connection with prototypes, which is *not* the same notion discussed below. And so I when I talk about lexical features in subsequent chapters I want it to be crystal clear what I am referring to.

2.3. On Chomskyan words—a closer look

So what then is a word of natural language? As indicated above, Chomsky describes words as *lexical items*, or again *LIs* for short. I have characterized Chomskyan words more specifically as particular instantiations of the triple <PHON, SYN, SEM>,

which is merely to say related bundles or unordered sets of phonological, morphosyntactic, and semantic *features*, where lexical features are taken to be "the elementary constituents of LIs and expressions constructed from them." So understood, we can think of each lexical feature as constituting a *partial description* of the words/LIs that it composes (where presumably some/many of the same features can be instantiated in multiple words/LIs). Thus, to understand the nature of Chomskyan words/LIs requires an understanding of the lexical features that comprise them.

Appeal to features has a long tradition in the cognitive sciences as a way of specifying, for example, the decompositional structure of lexical concepts, lexical meanings, and/or the definitions (or prototypes) of corresponding concept-*words* (which is not to be confused with what I am calling concepts$_W$). In these contexts the notion of a "semantic feature," or what are otherwise known as "semantic primitives," are usually cashed out in terms of basic ontological categories such as CAUSE, EVENT, PHYSICAL-OBJECT, and so forth. I will revisit this conception of features in the next chapter. But for the non-specialist it is important to stress immediately that this is *not* how Chomsky employs the technical term "feature" with respect to I-languages. Rather, linguistic features are posited as primitive mental representations that encode *purely formal*, *intrinsic* properties of words—i.e., properties that bear no direct relationships to entities in the mind-external world.

Understood in this way, Chomskyan "words" themselves are not what one might pre-theoretically conceive them to be. Yet this should come as little surprise. For there are compelling reasons to believe that like other dedicated, domain-specific faculties of the human mind such as early vision, the internal resources and representations of I-

languages, including their stock of lexical primitives, are by and large *cognitively impenetrable*, which is to say *non-conceptual* and thus *sub-personal*, *except* perhaps as manifest at the interfaces to language-external systems. It is of course undeniable that we consciously use/utter words to talk about things in the world. And, to be sure, the overt manifestation of words in speech provides evidence of their psycholinguistic reality and, moreover, their underlying nature.[20] The point, however, is that if words just are items in a Chomskyan lexicon we should not expect to introspectively recognize them *as such*, any more than we should expect to be able to introspect other *abstracta* such as edges, contours, and vertices as represented (i.e., instantiated) by the early visual system.

For those familiar with the view, it is also true that Chomsky ascribes intuitively descriptive names to linguistic features corresponding to their external *effects* in language use. For instance, Chomsky hypothesizes that common nouns encode binary features such as [±Human], [±Animate], and [±Artifact]. However, any relation between the computational properties of such features and the worldly entities they name is entirely abstract and thus *non*-constitutive of their informational contents. To repeat an earlier point, I-languages are posited to be internal systems of human cognition and only contingently useful for referring to things in the world. Rather, strictly speaking Chomskyan often stresses that linguistic features are posited to be purely syntactic in nature in the sense of being primitive mental symbols that participate in language-internal computational processes.[21] For example, what I am calling SYN features are conceived as

[20] This point is crucial to the relation between words and both our ordinary and scientific conceptions of them, which I explain in greater detail in Chapter 3.

[21] As Chomsky (2000a: 125) puts it: "The elements of these symbolic objects [i.e., LIs] can be called "phonetic" and "semantic" features, respectively, but we should bear in mind that all of this is pure syntax and completely internalist. It is the study of mental representations and computations, much like the inquiry into how the image of a cube rotating in space is determined from retinal stimulations, or imagined."

intra-system "instructions" between combinatorial processes/operations for how to build interpretable syntactic structures. By contrast, phonological features, or again PHONs in the Chomskyan vernacular, are so-named because they have certain *consequences* with respect to their interpretations at the interface to A-P. More specifically, with respect to speech production Chomsky's describes PHON features as instructions to the articulators for how to pronounce words (usually in combination with other words). For instance, given the capacity to pronounce its constituent phonemes, the PHON features of 'dog' are understood by A-P, collectively, as an instruction for how to pronounce the word 'dog'. In turn the capacity to pronounce the string 'brown dog' depends on knowing how to pronounce 'brown' and 'dog' yet which may be articulated slightly differently when uttered in combination in fluent speech.

If still a bit unclear, Chomsky puts it this way: [22]

> To say that phonetic features are "instructions" to sensorimotor systems at the interface is not to say that they have the form "Move the tongue in such-and-such a way" or "Perform such-and-such analysis of signals." Rather, it expresses the hypothesis that the features provide information in the form required for the sensorimotor systems to function in language-independent ways. Similar observations hold on the (far more obscure) meaning side.

In regard to this last remark, SEM features are said to have semantic import with respect to their interpretations at the interface to C-I. I will say more about this below, but in

[22] Chomsky (2000b: 91) writes "To say that phonetic features are "instructions" to sensorimotor systems at the interface is not to say that they have the form "Move the tongue in such-and-such a way" or "Perform such-and-such analysis of signals." Rather, it expresses the hypothesis that the features provide information in the form required for the sensorimotor systems to function in language-independent ways. Similar observations hold on the (far more obscure) meaning side." More specifically, Chomsky (2000a: 36) states that "…a lexical item provides us with a certain range of perspectives for viewing what we take to be the things in the world, or what we conceive in other ways; these items are like filters or lenses, providing ways of looking at things and thinking about the products of our minds. The terms themselves do not refer, at least if the term refer is used in its natural language sense; but people can use them to refer to things, viewing them from particular points of view—which are remote from the standpoint of the natural sciences, as noted." In short, Chomsky states (*ibid*: 34): "We might well term all of this *a form of syntax*, that is, the study of the symbolic systems of C–R theories ("mental representation")." [my emphasis].

terms of comprehension we can again think of SEMs as being understood by "belief systems" (at the C-I interface) as *instructions to activate semantically related extralinguistic concepts*. In general, lexical features impose selectional restrictions on which words can combine with which others in grammatically permissible ways, and, with a little luck, conceptually coherent ways. More specifically, Chomsky (2000a: 36) adds that:

> […] a lexical item provides us with a certain range of perspectives for viewing what we take to be the things in the world, or what we conceive in other ways; these items are like filters or lenses, providing ways of looking at things and thinking about the products of our minds. The terms themselves do not refer, at least if the term refer is used in its natural language sense; but people can use them to refer to things, viewing them from particular points of view—which are remote from the standpoint of the natural sciences, as noted.

The upshot is that "We might well term all of this *a form of syntax*, that is, the study of the symbolic systems of C–R theories ("mental representation");" (*ibid*: 34, my emphasis).

The general moral of this discussion is that one should not draw conclusions from theoretical naming conventions about the ontological status of linguistic features (e.g., [±Human], [±Animate], and [±Artifact]). Rather, from an I-language perspective lexical features were initially posited to explain certain facts/generalizations about the phonological, morphosyntactic, and semantic behavior of words with respect to language-*internal* computational processes *as reflected* in their interpretations at the interfaces to language-external cognitive systems. This is simply to stress that on Chomsky's view I-languages are primarily systems of *thought* and only contingently useful for the externalization of those thoughts in interpersonal communication. By contrast, from an

externalist/E-language perspective the most we can say about the intrinsic semantic properties of Chomskyan words is that they guide and constrain without determining what those words can coherently be used/uttered to talk about in ordinary conversational situations.

2.3.1. A closer look at feature theory

From an historical perspective, the theory of "distinctive features" (as they are sometimes called) was first introduced by linguists in the 1950's as a way to classify the basic units of phonological structure called *segments* or *phonemes*. More specifically, phonological features were employed to characterize the internal representation of a word's syllabic and prosodic structure. Traditionally, phonological features, or again PHONs in Chomsky's terminology, are specified using binary feature "markings" such as [±Consonantal], [±Sonorant], [±Nasal], and [±Syllabic], where a positive specification indicates the presence of the named feature/property and a negative marking indicates its absence.[23] Contemporary phonological feature systems are much more sophisticated than this, however. But roughly speaking, phonological features are again posited to capture certain facts/generalizations about how the internally represented phonological structures of words determine their permissible phonetic combinations and overt pronunciations in spoken language. In terms of comprehension, these features are used by the parser to segment linear speech signals into the discrete structural units (i.e., *words* or *morphemes* roughly speaking) that participate in language-internal computational processes.

As early as 1965 Chomsky had co-opted the notion of lexically-specified features in lieu of production/rewrite rules to describe (and ultimately explain) observed restrictions

[23] Other proposals allow features to be single-valued, or to have more than two values, thus dispensing with the binary notation. But for ease of exposition I'll stick with the binary notation.

on the syntactic distributions of words—i.e., the linguistic contexts into which various lexical categories can be inserted (their so-called "insertion frames") while preserving grammaticality. We know, for example, that in English only verbs and prepositions take nouns (or noun phrases/NPs), as direct complements, and that only nouns/NPs and adjectives/APs can be the complements of determiners. Generally speaking, each major lexical category stands in complementary distribution to the others.[24] And it appears that the grammatical rules that enforce such restrictions must be capable of distinguishing the various grammatical categories of words. Such observations are what motivated Chomsky to posit feature markings such as [+N] for common nouns, [+V] for verbs, [+A] for adjectives (or adpositions, more generally), and so on, to identify words/LIs by their major grammatical category.[25] In addition, Chomsky (1965) introduced "subcategorization" and "selectional" features to capture more fine-grained distributional patterns among various subclasses of words, such as count vs. mass nouns and transitive vs. intransitive verbs.

More will be said below about the classification of lexical features and the restrictions they impose. But again for the benefit of non-specialists, perhaps the easiest way to understand the theoretical role of lexical features is by way of analogy. In particular, we might compare Chomskyan words/LIs to pieces of jigsaw puzzle whose

[24] The major open-class lexical categories correspond loosely to the traditional "parts of speech," that standardly include nominals (i.e., nouns, both common and proper), verbs, adjectives, and on some views both prepositions and adverbs. However, Chomsky sometimes speaks of adverbs as a lexical subcategory. In any case, he continues to consider their category labels to be a defining feature of Words. For example, he writes (1995:30): "…any theory of language must include some sort of lexicon, the repository of all (idiosyncratic) properties of particular lexical items. These properties include a representation of the phonological form of each item, *a specification of its syntactic category*, and its semantic characteristics." [my emphasis].

[25] Chomsky (1995: 30) states that among the properties of the lexicon "include a representation of the phonological form of each item, a specification of its syntactic category, and its semantic characteristics." However, more recently Chomsky (2006) has floated the possibility that phrase structures are "bare," which is to say they lack category labels, even at the terminal level.

physical shapes are determined by certain intrinsic properties of their types. Conversely, the shape of each piece constrains its possible position within the puzzle and thus the overall shape/structure of the puzzle itself. By analogy, the intrinsic properties of words/LIs represented by lexical features restrict their combinatorial potential—i.e., their possible insertion frames within the overall structure of larger expressions in which they occur. Words that belong to the same grammatical category share many of the same properties and thus "fit" in many of the same places. Likewise, common categories "select for" the same number and type of grammatical complements. More specifically, according to Chomsky phrasal structures are *projections* of the formal features of their syntactic heads[26]—a thesis known as *Projectionism*, which typically goes hand-in-hand with what above I called *Lexicalism*.[27] On this count Chomsky (1986: 81) writes:

> […] phrases typically consist of a head (noun, verb, adjective, preposition, and possibly others) and an array of complements determined by lexical properties of the head. The category consisting of the head and its complements is a *projection* of the head (NP if the head is an N, VP if the head is a V, etc.).

Without yet saying more precisely what lexical features are, one first wants to know *to what*, if anything, such features attach? That is, linguists often speak of feature "markings," and so one wants to know what it is, exactly, that is being marked? At first blush, we might take the structural bearers of lexical features to be *grammatical formatives*—i.e., the linguistic "objects" that are thought to be manipulated by language-

[26] As Radford (2009) defines the notion, "The head (constituent) of a phrase is the key word which determines the properties of the phrase. So, in a phrase such as 'fond of fast food', the head of the phrase is the adjective *fond*, and consequently the phrase is an adjectival phrase (and hence can occupy typical positions associated with adjectival expressions—e.g. as the complement of *is* in 'He is *fond of fast food*'). In many cases, the term *head* is more or less equivalent to the term *word*…"

[27] Variants of the Projectionist approach to morpho-syntax can be found in Chomsky (1970), Grimshaw (1979), Pesetsky (1982), Hale & Keyser (1993, 2002), Baker (2003), and Levin & Rappaport Hovav (1998, 2005).

internal operations.[28] A prior question, then, is what these formatives are supposed to be? Let me briefly reconstruct a commonly assumed answer to this question followed by my best interpretation of Chomsky's actual view on the matter, which I sense is sometimes misunderstood.[29]

2.3.2. On morphology (word structure)

In the domain of phonology, we can begin with the commonsense observation that many if not most words of natural language are phonologically complex, which is to say *polysyllabic*.[30] More specifically, speech signals (utterances/vocalizations) are parsed by the phonological system into abstract structural segments, again called *phonemes*, which are defined as the smallest pronounceable units of sound. Phonologically complex words are therefore *polyphonemic*—their language-internal representations consist in roughly word-sized sequences of phonemes that linguists sometimes call *phonological words* and whose internal phonemic structures are sometimes referred to as their phonological "spellings."[31] Phonological *features* are in turn said to determine a word's phonemic structure/spelling. So understood, one is given to assume that individual *phonemes* are the structural bearers of phonological features.

---

[28] (Chomsky [1965: 86]) writes, for example, that "…separating the lexicon from the system of re-writing rules has quite a number of advantages. For one thing, many of the grammatical properties of formatives can now be specified directly in the lexicon, by association of syntactic features with lexical formatives, and thus need not be represented in the re-writing rules at all."

[29] Indeed, Chomsky is sometimes unclear about these details, and his terminology is occasionally inconsistent. And so I'm not sure that I fully understand the view. But again what follows represents my best reconstruction of it.

[30] Some words also have complex stress and/or intonation patterns, but we can safely gloss past these details for purposes here.

[31] To be clear, I am using the term "spelling" here as a metaphor for the *abstract phonemic structure* of structurally complex phonological words, which I assume to be distinct from (i) the physical acoustic structures of their token-utterances, which have *phonetic spellings*, and (ii) their physical inscriptions, which have *orthographic/alphabetic spellings*, both of which are mind-external notions.

In the domains of morphology and syntax, *morphemes* are standardly taken to be the smallest syntactically manipulable units of natural language, related sequences of which constitute what some linguists refer to as *grammatical* or *syntactic words*.[32] However, when Chomsky and colleagues speak of words they are quite often referring specifically to primitive syntactic formatives known as *lexical roots*—morphological units that may consist of multiple morphemes yet which occupy the terminal nodes of syntactic trees.[33] Chomsky (1995: 20) states, for example, that "The lexicon contains the root [walk], with its idiosyncratic properties of sound, meaning, and form specified."[34] Characterized in this way, one might assume that lexical roots are the structural bearers of lexical features. However, Chomsky (2000a: 170) also states that "there is no separate substratum, *the word*, in which the properties [features of LIs] inhere." Now, as I read him Chomsky is not saying that words as such do not exist. Rather, I take him to mean that words are nothing over and above the set of lexical features that comprise them. On this construal, in other words, particular instantiations of the triplet <PHON, SYN, SEM>

[32] However, as Julien (2006: 619) observes "Using the term 'grammatical word'… is not strictly correct because phonology is, of course, also a part of grammar."

[33] Indeed, what one might count as words of certain polysynthetic languages can be massively polymorphemic.

[34] Now, it's not entirely clear what Chomsky means here by "form." Though what he is surely *not* referring to is the word's *orthographic form* in written language (i.e., its alphabetic spelling). For contrary to what some theorists seem to think, I am aware of neither statements to the effect, nor evidence to suggest, that Chomskyan LIs encode the conventional spellings of words. This should come as no surprise, as language learners learn to speak long before they learn to write (if ever). Indeed, written language is presumably nothing more than the transcription of phonetic speech signals into their customary written/orthographic forms according to certain conventionalized rules. And while literate speakers eventually learn these rules, they again presumably do not represent the orthographic forms of words. Rather, by hypothesis what LIs represent are abstractions over word-*sounds*—i.e., phonetic speech signals—whose "image" (as Chomsky sometimes puts it) is encoded by the PHON component of their lexical entries. Thus, from a psycholinguistic perspective the theoretically prior question is how speech signals are internally parsed, represented, and interpreted by I-languages; a question to which I return below. The upshot is that like other basic grammatical formatives such as *phonemes* in the phonological domain, lexical roots are again by hypothesis *mental representations*—i.e., complex mental *symbols* whose tokens enter into certain language-internal computational processes. As such, one assumes that lexical roots are distinguished (i.e., type-individuated) by the physical "shapes" of their tokens (presumably as realized by certain brain states). We might then think of the internal "shape" of a lexical root as its *morphological form* (which is again not to be confused with its conventional orthographic form or its phonetic realization in spoken language).

just are the linguistic objects/formatives manipulated by language-internal computational processes.

Chomsky adds that "LIs may be decomposed and reconstructed in the course of computation" and, moreover, that "any feature change yields a different LI." These last two remarks in isolation are also a bit misleading, however, and thus require further unpacking. It will help to begin here by briefly reviewing the distinction between "open-class" content words and "closed-class" functional items. Most mainstream versions of Minimalist syntax count as members of the open-class lexicon the categories (N)oun, (V)erb, and (A)djective—lexical categories whose members usually have inflected and/or derived forms, independently specifiable meanings, and the set of which can in principle grow without limit as speakers acquire new words (hence the label "open" class). Correspondingly, these items head the phrasal categories NP, VP, and AP, respectively.[35]

The open-class lexicon is to be contrasted with a comparatively small and largely static inventory of closed-class functional items whose respective roles in the language are more or less fixed by the grammar. The closed-class lexicon standardly includes sentential connectives/copulas, complementizers, auxiliaries, pronouns, determiners, quantifiers, and certain inflectional morphemes. These items in turn head their own phrasal categories such as CP (complementizer phrase), TP (tense phrase), and DP (determiner phrase). Unlike open-class categories, however, functional heads typically do not themselves have inflected or derived forms. Indeed, not all functional elements are realized in the syntax as lexical roots, *per se*, which is to say independently pronounceable lexical items. In particular, elements such as T(ense), Asp(ect), and

[35] While controversial, some theorists treat Adv(erbs) and P(repositions) as open-class items, while most take them to be closed-class elements, as discussed next. However, this question is largely irrelevant to my concerns here.

Agr(eement), which head the abstract functional categories TP, AspP, and AgrP, are realized, syntactically, as bound morphemes that are spelled-out, phonologically, as affixes on open-class roots. For example, the suffix '-*s*' that attaches to English verbal roots signals agreement with third-person singular NP-subjects (i.e., subjects marked [+3rd Person], [+Singular]), as in "He walk*s*"). However, other posited functional categories have no phonological realizations but are merely needed, as it were, to keep the grammar happy—they are home to certain features that determine relevant structural relations between open-class categories. But setting aside these details, the takeaway here is that closed-class lexical items constitute the functional backbone of syntactic structures; i.e., the structural glue that holds together their open-class lexical constituents.

More generally, the proposal is that the grammar operates on what Chomsky calls a *Numeration*—an ordered list of open-class LIs along with relevant closed-class functional elements selected from the lexicon that serve as the raw materials for combinatorial/derivational processes. The computational procedure then recursively combines these items according to their featural specifications and the rules of grammar to generate more complex, hierarchically structured linguistic expressions (i.e., structured bundles of linguistic features). Once introduced, certain features can be individually manipulated by the syntax in the course of a derivation.[36] For example, phrasal heads can inherit certain features from and/or assign features to subtending LIs. Generally speaking, individual features can be moved, copied, reassigned, and in some cases deleted in order

[36] Chomsky's *Inclusiveness Condition* prohibits new features/LIs from being introduced into the Numeration during the course of computation, as explained in Chomsky (2000b: 100): "UG makes available a set *F* of features (linguistic properties) and operations *CHL* (the computational procedure for human language) that access F to generate expressions. The language L maps F to a particular set of expressions Exp. Operative complexity is reduced if L makes a one-time selection of a subset [F] of F, dispensing with further access to F. It is reduced further if L includes a one-time operation that assembles elements of [F] into a lexicon Lex, with no new assembly as computation proceeds."

to satisfy legibility conditions imposed at the interfaces to language-external consumer systems. The upshot, according to Chomsky (2000a: 175), is that "At these interface levels there may be no sub-unit corresponding to LI."

Now, on the one hand the picture that emerges from this (all too brief) discussion is a conception of words as transient, dynamically reconfigurable bundles of lexical features. On the other hand, Chomsky's notion of the lexicon as a stable compendium of sound-meaning pairs implies a conception of words as consisting in a fixed set of lexically-encoded features that persist over time. That is, the persistence of at least some lexical features is presumably what allows speakers to recognize and utter tokens of the same word on different occasions. Thus, *I think* the picture that Chomsky has in mind is that tokens of the same word—i.e., *the same lexical root* "with its idiosyncratic properties of sound, meaning, and form specified"—can assume different *morphological forms* by combining with various functional morphemes/features in different linguistic contexts to generate distinct word-forms having distinct lexical features. Otherwise, one assumes that the lexical roots themselves are type-individuated by the context-*invariant* features encoded by their lexical entries. Chomsky does in fact suggest elsewhere (see notes 19 and 25) that certain type-individuating features of words are lexicalized as part of the word-acquisition process. Or anyway, I will be assuming as much in what follows.

Let me pause here briefly to restate the point of this discussion, which is an attempt to develop a conspicuous conception of what words of natural language are, and which under the present hypothesis are the primitive mental representations that constitute part of the stable state of a competent speaker's mind (or brain) known as an I-language. But more than this, the acquisition and representation of words constitute part of a speaker's

core linguistic competence. In explaining the nature of such competence I have made reference to certain aspects of linguistic processing that count as what Chomsky refers to as *performance*—the various ways in which language is put to use in the service of communication. However, while certain aspects of linguistic processing/performance figure into my core thesis, I want to make clear before proceeding that this is no part of Chomsky's agenda.

So qualified, let me also continue to delay a more detailed discussion of lexical features in favor of trying to first get a bit clearer about how words are distinguished as such by language-internal processes. But in the meantime we can again think of lexical features as partial descriptions of the words/LIs they compose. It is also widely assumed that the inventory of such features is finite, universal, and by definition represented in a speaker's long-term lexical memory. Now, if what it means to be an LI is simply to be listed in the lexicon as such, then strictly speaking individual lexical features are LIs. However, Chomsky again speaks of *words* (both open- and closed-class) as *consisting in* related *bundles* of lexical features. The relevant consequence is that, again strictly speaking, not all LIs are words. And if certain assumptions below are correct, then not all words are LIs. The more pressing question, however, is how I-languages distinguish words from bound morphemes, on the one hand, and more complex expressions (i.e., phrases and sentences) on the other. In what follows, I will again offer my best interpretation of Chomsky's view on the matter.

2.4. On the nature of Words—to a second approximation

To a second approximation, I have characterized words of natural language (both open and closed-class) as *lexical roots*—i.e., basic syntactic formatives composed of

bundles of lexical features (i.e., instances of the triplet <PHON, SYN, SEM>) that participate in language-internal computational processes and that occupy the terminal nodes of syntactic trees. Lexical roots in turn combine, syntactically, with bound morphemes and perhaps other functional elements to yield inflected and/or derived word-forms that are composed of, or generated from, *multiple* LIs. Some such constructions, and in particular non-productive (irregular/suppleted/declensional)[37] word-forms such as the English past tense verb 'went', the comparative adjective 'better', its superlative 'best', and so forth, are also standardly taken to be listed as such in the lexicon.[38] By contrast, fully productive (i.e., regular/rule-governed) word-forms could in-principle be computed online according to morphological rules each time they are uttered and/or interpreted. However, many researchers hypothesize that in the interest of computational economy certain high-frequency yet fully productive/regular word-forms may also be recorded somewhere in long-term lexical memory.[39]

Further evidence suggests that other complex "word-like" constructions formed initially through morphological processes such as compounding (e.g., 'egghead') and incorporation (e.g., 'babysitter') are retrieved from lexical memory as pre-built "chunks" and thus behave like words with respect to the grammar.[40] Rather than being recorded in the lexicon, however, these latter constructions—particularly those with idiosyncratic/idiomatic meanings—are sometimes said to be separately listed as part of a

[37] Suppletion is a variety of allomorphy that results in the substitution of one phonological form with another. Declension is a form of inflection that distinguishes lexical categories according to Number, Case, and Gender (e.g., the singular noun 'goose' vs. its plural 'geese').
[38] Though many such forms are often not recognized as such by lexicographers.
[39] Wunderlich (2006: 9) argues, for example, that "words can be morphologically complex and semantically transparent, so that they do not belong to the idiosyncratic knowledge, but if they are frequently used they can nevertheless be memorized and thus belong to the 'mental lexicon'."
[40] See Libben & Jarema (2007) for discussion.

speaker's mental "vocabulary."[41] Setting details aside, the relevant consequence is again that not all LIs are properly classifiable as words. And if vocabulary items count as words, then strictly speaking not all words are classifiable as LIs.

To consolidate on some terminology while at the same time avoiding certain unintended theoretical commitments, let me introduce the technical term '*Word*' to designate pronounceable bare/uninflected lexical roots that (i) minimally encode (are constituted by) features/properties of sound and meaning, (ii) are listed as such in a speaker's mental lexicon, and (iii) occur as terminal nodes in syntactic trees.[42] As discussed further below, it may also be a defining property of Words to be marked for major grammatical category.[43] But for present purposes it's enough to assume that most Words are represented by (or within) I-languages as bare lexical roots whose tokens are again instantiations of the triplet <PHON, SYN, SEM>. Broadly speaking, however, my notion of a Word is intended to cover any linguistic constituent that satisfies the conditions above and is (or potentially can be) manipulated by phrasal syntax as an atomic (indivisible) lexical unit, including but not limited to non-productive compounds, incorporations, and portmanteaus (e.g., 'smog').[44] By contrast, I assume that Words differ from bound morphemes and clitics in that the former have independently specifiable meanings and are individually pronounceable with their meanings.[45]

---

[41] Notice that this is not the sense of 'vocabulary' as used by Distributed Morphologists.

[42] For those who care about such questions, I'm willing to grant that representations which constitute what I am calling the mental lexicon are distributed across different brain regions, as is commonly assumed by PDP models in the field of psycholinguistics.

[43] *Cf*., Chomsky (1965), Di Sciullo & Williams (1987).

[44] This is essentially what Di Sciullo & Williams (1987) dubbed the principle of "Lexical Integrity." However, Booij (2007: Ch.12) suggests that the principle of Lexical Integrity is perhaps too strong, as there appear to be counterexamples.

[45] It is sometimes claimed that some bound morphemes have no independently specifiable meanings. However, emphasis on the conjunct here is to recognize that some bound morphemes, and perhaps also

2.4.1. Accommodating some facts

Chomsky and his followers again posit lexical features to explain a range of facts/generalizations about the phonological, syntactic, and semantic behavior of Words/LIs with respect to their grammatical distributions and interpretations. Below I survey a small but relevant selection of such facts along with some suggestions about how to accommodate them within a Chomskyan framework for natural language. This discussion will be useful, in particular, as segue into the central goal of this chapter, which is to develop an empirically plausible hypothesis about the nature of *lexical meanings*. However, several of the more general facts about words/language mentioned here are in one way or another relevant to a deeper understanding of my core thesis, many of which will reappear in later chapters.

2.4.1.1. Some facts about phonology

On the side of phonology, one robust generalization is that the relation between physical speech signals and their internal phonological representations is *many-to-one*. Consider, for example, that speakers have unique accents (voice quality, roughly speaking) and often speak different regional dialects (e.g., you say /temaytoh/, I say /tomahtoh/). For that matter, consider a native English speaker trying to parse the broken English of his/her immigrant parents. In either case, competent listeners usually have little difficulty (within limits) recognizing phonetically distinct utterances as utterances of the same Word.[46] Such facts suggest that I-languages are capable of extracting and

---

clitics, do appear to have independently specifiable meanings. For example, the English derivational suffix '-er' appended to any verb, *V*, robustly means something like "one who *V*s; e.g., 'dancer', 'thinker', etc.

[46] As Dahan & Magnuson (2006:252) write: "When someone speaks, the linguistic content and speaker characteristics (e.g., physiology of the vocal tract, gender, regional origin, emotions, identity) simultaneously influence the acoustics of the resulting spoken output. Additional sources of variability

registering a range of acoustically distinct lexical speech signals as tokens of the same abstract phonological word, again understood as a sequence of phonemes. Such facts are again accounted for, in part, by positing corresponding phonological features (PHONs) encoded by the Word's lexical entry. It is important to stress that PHONs are *abstractions* over lexical speech signals in the sense that while neither structurally or physically isomorphic, they must at some level of analysis be inter-transcribable.[47]

In short, it seems clear that each PHON subsumes a wide (yet limited) range of acoustically distinct speech signals. This observation raises the important empirical question of just how much variation can be tolerated for purposes of comprehension, and thus successful communication.[48] From a metaphysical perspective, this question is important because the answer bears directly on the individuation of Words as they are represented in the lexicon. Yet to my knowledge, no one has a definitive answer to this question. However, I trust alongside the experts that there is a fact of the matter. From a computational perspective, I suspect that the shape of the explanation will be similar to an explanation for how the visual system tracks the constancy of surface colors across variation in texture, lighting conditions, viewer perspective, and so forth (*cf.*, McGilvray [1994] for discussion).

---

include rate of elocution, prosodic prominence, and the phonetic context in which each word is pronounced. Nonetheless, listeners are able to recognize acoustically different stimuli as instances of the same word, thus extracting the similarity that exists between these different tokens, and perceiving them as members of the same category."

[47] Chomsky (1986: 57) says, for example, that "… lexical items can be given in an abstract form in phrase structure representation, then converted by a succession of phonological and phonetic rules to their actual phonetic form…"

[48] Of course, the *culturally acceptable* range of variation is often artificially determined by prescriptive standards of correctness designed to ensure conformity; that is, to constrain variation so that for various purposes it does not impede everyday communication. However, Kaplan (1990: 101) contends that, technically speaking, "the difference in phonographology, the difference in sound or shape or spelling, can be just about as great as you would like it to be."

In short, whatever is the correct explanation with respect to speech processing, the boundaries are clearly both vague and subject-dependent, yet also subject to change/improvement over time and with experience.[49] Such observations illustrate the fact that the informational contents of PHONs bear no *direct* correspondence to their physical manifestations in speech. This is of course *not* to say that there is no relation whatsoever between PHONs and their physical manifestations. Rather, Chomsky's view is that the relation is at best *causally determinative* as opposed to *constitutive*. As we will see below, the same goes for SEM features in the semantic domain.

2.4.1.2. Some facts about syntax

On the side of syntax, lexical features are again posited to explain observed contextual restrictions on the possible syntactic distributions of various lexical categories—i.e., restrictions regarding which categories can combine with which others in various linguistic contexts while preserving grammaticality (i.e., without violating *syntactic* rules as opposed to *semantic/conceptual* constraints). As mentioned earlier, for example, only verbs and prepositions take nouns/NPs as direct complements (or arguments), which explains why (1) and (3) below are grammatically acceptable while (2) and (4) are anomalous:

(1) Max broke the window

(2) *Max the window broke

(3) Max declared his innocence to the deed

(4) *Max declared to his innocence the deed

[49] For example, I assume that most of us have learned to hear/understand the Word-sound \havad\ as pronounced by native Bostonians to mean Harvard University.

In the early days of generative grammar, such facts were captured by lexical production rules as part of the base set of phrase structure rules. These rules again make explicit reference to lexical and phrasal categories. As such, Chomsky (1965) hypothesized that Words encode feature-markings such as [+N], [+V], [+A], and so forth corresponding to a Word's major grammatical category.

However, it was apparent then, as it is now, that most if not all open-class lexical categories pattern into more fine-grained subcategories according to the exact number and types of their grammatical complements. For instance, common nouns—Words lexically marked [+N]—subcategorize as either count or mass, prompting Chomsky's introduction of the selectional feature [±Count], where a negative marking here indicates a mass term. For example, the lexical entry for the English count noun 'chair' includes the feature specification [+N, +Count], whereas inherently mass terms such as 'furniture', 'mud', etc., are marked [+N, -Count]. Among other constraints, the [±Count] feature is invoked to capture the fact that mass terms do not naturally combine with plural morphemes (e.g., *'furnitures', *'muds'), or doing so forces a count reading (e.g., using 'wines' to mean *bottles* of wine). Mass terms also cannot combine directly with the determiners 'a'/'an', or quantifiers such as 'five', 'each', 'few', and 'many'. In the reverse direction, determiners and quantifiers differentially select for NP complements whose head noun carries one of the two feature-markings [+Count] or [-Count]. For instance, the lexical entry for 'every' must specify that it takes only NP complements whose head is marked [+Count].

Common nouns subcategorize in other ways as well. In general, Chomsky (1965) analyzes the class of nominal expressions, N, according to the following subcategorization hierarchy:

(i) N → [+N, ±Common]

(ii) [+Common] → [±Count]

(iii) [+Count] → [±-Animate]

(iv) [-Common] → [±Animate]

(v) [+Animate] → [±Human]

(vi) [-Count] → [±Abstract]

By way of a full example, Chomsky proposes that the lexical entry for the Word 'boy' (i.e., the lexical root √*boy*) consists in the following selectional features:

√*boy*: [+N, +Common, +Count, +Human]

Notice that there is no need here to positively specify the feature [+Animate], as this follows from the feature-marking [+Human]. Strictly speaking, Chomsky goes so far as to suggest that the category label [+N] is itself predictable from the other specified features and is therefore otiose. Indeed, this latter proposal has recently reappeared as part of what Chomsky (2005b: 14) calls "bare phrase structure." However, as reported by Radford (2009: 77) this hypothesis has yet to garner widespread support from the broader Minimalist community.

Yet under the plausible assumption that 'boy' is lexically marked as a count noun, it can legitimately combine with the definite determiner 'the' to form the determiner phrase/DP 'the boy'. Conversely, Chomsky (1965) hypothesizes that the lexical entry for

'the' carries a subcategorization feature that licenses its combination with count nouns such as 'boy'. Specifically, this selectional restriction can be specified as follows,

√*the*: [+D, +Definite, — [+N] — [+Count], …]

where collectively the categorial feature [+D] and the subcategorization feature [+Definite] identify the *functional* root √*the* as a definite determiner, and the selectional feature '— [+N] — [+Count]' indicates that the Word 'the' "selects for" a count noun as one of its direct complements (the ellipsis '…' indicating that it selects for other categories as well).

Verbs, by contrast, subcategorize for both the number (i.e., valence/adicity) and type of their internal grammatical arguments (i.e., direct and indirect objects). For example, Chomsky proposes that the lexical entry for a verb like 'frighten' can be specified as follows,

√*frighten*: [+V, — NP, — [+N] — [-Animate], …]

where here the first selectional feature '— NP' indicates that 'frighten' takes an NP as its direct object, whereas the second feature [— [+N] — [-Animate]] indicates that the head noun of that NP cannot designate an inanimate object (i.e., it must refer to a sentient being).[50] By contrast, the lexical entry for verbs such as 'eat' and 'grow' encode, in addition to [+ — NP] the feature [— #] reflecting their status as verbs that permit direct object omission/deletion. In short, the feature theory introduced by Chomsky (1965) implements a kind of agreement system where the subcategorization features of Words must agree with the selectional features of their complements, and *vice versa*.

[50] As a point of historical interest, Chomsky (1965) flirts with the subcategory feature [+/-Transitive] to distinguish transitive from intransitive verbs. But here again the adicity of 'frighten' follows from the fact that only a direct object is specified by its lexical entry.

I haven't the space here for a more comprehensive review of Chomsky's (1965) feature system, which is fairly sophisticated in its finer details.[51] Rather, this discussion is merely intended to give a "feel" for the theoretical role that lexical features were originally invoked to play, and indeed continue to play in contemporary generative linguistics. However, to round out the present example, notice how the feature specifications of 'boy' and 'frighten' capture the fact that while (5) below sounds fine, (6) is somehow anomalous:

(5) Sincerity may frighten the boy

(6) *Sincerity may frighten justice

The postulated reason for the anomaly in (6) is again that so-called "psych" verbs such as 'frighten' only select for NP complements marked [+Animate]. By contrast, the noun 'justice' presumably subcategorizes as [+Abstract], which implies both [-Human] and [-Animate]. This mismatch in selectional/subcategorization features thus putatively explains the infelicity of (6).

Now, it is a point of contention whether (6) is *grammatically* anomalous as opposed to merely invoking a thought that is *conceptually* odd or otherwise incoherent. That is, the question is whether the unacceptability of (6) owes to a language-*internal* syntactic violation or a language-*external* conceptual constraint. To flag (6) with a '*', as many linguists would, is to treat it as ungrammatical. But to my mind the anomaly of (6), unlike say (4) from above, is on a par with sentences like "Colorless green ideas sleep furiously," which while nonsensical Chomsky considers to be perfectly well-formed, grammatically speaking (and I happen to agree). But setting individual cases aside, this

[51] And so far as I can tell, much of what Chomsky said in 1965 is still considered valid.

distinction is important because it bears on the division of lexical labor introduced earlier between SYN and SEM features, and is therefore worth dwelling on a moment longer.

2.4.1.3. Distinguishing SYNs from SEMs

First of all, one can agree that the infelicity of (6) is attributable to a feature mismatch without assuming that it violates a rule of *grammar*. More specifically, the suggestion I have in mind goes hand-in-hand with Chomsky's more recent distinction between "c-selection" features (for *category* or *constituent* selection) and what he calls "s-selection" features (for *semantic* selection), which corresponds roughly to the distinction I am making between SYNs and SEMs. As I understand the distinction, the subclass of c-selection features, which subsumes both major category and subcategorization features, are purely formal, language-internal devices employed by I-languages solely in the service of building interpretable syntactic structures.[52] More specifically, c-selection features impose constraints on the combinatorial potential of words, violations of which result in *ungrammaticality*, which is to say that they lead to a "crash" (or lack of "convergence") in the derivational process.

By contrast, s-selection features are by hypothesis visible to and thus directly interpretable by conceptual consumers (or again "belief-systems" as they are sometimes called) at the C-I interface. In other words, while also grammatically relevant, s-selection features have *direct consequences* for the semantic interpretations of their host Words (relative to a context of utterance). I'll say more below about what those consequences are, but we can think of the s-selection features of a Word, collectively, as constituting its *linguistic meaning*. But for the sake of clarity, I will hereafter abandon the notions of "c-

[52] Or to borrow an apt metaphor from Paul Pietroski (p.c.), we can think of c-selection features as formal devices that are necessary "to keep the trains running on time."

selection" and "s-selection" in favor of the labels 'SYN' and 'SEM', respectively. Likewise, I will henceforth think of the SEM features of a Word, collectively, as constituting various aspects of its lexical *meaning*.

2.4.1.4. Some facts about lexical meanings (SEMs)

As with just about everything else in theoretical linguistics, the grammatical significance of SEMs/lexical meanings is a hotly debated issue. However, it is widely agreed that syntactic structures are projections of the lexical features of their constituents. In particular, certain SEM features of verbs are said to determine the syntactic realization of their argument structure. As Levin & Rappaport (2005: 7) report, for example:

> Since the 1980s, many theories of grammar have been built on the assumption that the syntactic realization of arguments—their category type and their grammatical function—is largely predictable from the *meaning* of their verbs. Such theories take many facets of the syntactic structure of a sentence to be *projections* of the lexical properties of its predicator—its verb or argument-taking lexical item." [my emphases].

Levin & Rappaport Hovav are referring specifically to Chomsky and others who embrace a theory of argument realization grounded in the *thematic relations* which hold between events and their event participants, where participant roles (or thematic–roles) are specified in terms of semantic notions such as *Agent*, *Patient*, *Experiencer*, *Goal*, *Path*, and so forth. As L&R report further (*ibid*: 8):

> Chomsky (1986: 86), for example, suggests that part of the semantic description of the verb *hit* is a specification that it selects arguments bearing the semantic roles agent and patient. He suggests that the syntactic type and grammatical relation of each argument (c-selection) can be derived from s-selection via general principles, so that subcategorization frames can be dispensed with altogether.

While the issues here are again controversial, the leading idea is that the participant roles of the events designated by various verbal subcategories correspond to the structural

positions of their grammatical arguments. For example, the grammatical subjects (or *external* arguments) of transitive action verbs always correspond to the thematic *Agent* of the designated event (roughly speaking, its *cause*), whereas their direct objects (or *internal* arguments) correspond to the event's *Patient*/*Theme* role (roughly, the thing *acted upon*).[53] Similar generalizations hold cross-linguistically for other postulated verbal subclasses.

I am deliberately keeping the discussion here general in the interest of both clarity and brevity. But of greatest import is Chomsky's *Projectionist Hypothesis*, which has become more or less orthodoxy in contemporary generative linguistics. The hypothesis, as I understand it, is that the syntactic realization of a verb's argument structure is a *projection* of its lexically-encoded thematic structure as mediated by "linking rules" that bind these two structures together. Importantly, the postulated lexical features that determine thematic relations are typically characterized as being *semantic* in nature, which is to say classifiable as SEM/s-selection features in Chomsky's terminology. In this sense, the SEM features of Words are said to be the "semantic determinants" of linguistic structure/form.

However, one might question whether the features that determine argument structure are in fact *semantic* in nature, which implies that they are directly interpretable at the interface to C-I. For one can imagine a combinatorial system that interprets SEM features, syntactically, as instructions to *introduce* into the emerging structure certain functional elements (e.g., little *v*) that *correspond to* semantic notions such as *Agents*, *Patients*, and so on. These functional elements are then interpreted by extralinguistic conceptual systems as something like instructions to activate the interpreter's *concepts*

[53] *Cf*., Williams (1994) for more on the distinction between internal and external arguments.

AGENT(x), PATIENT(x), and so forth; *cf*., Pietroski (2005). In other words, the suggestion here, which strikes me as quite plausible, is that what Chomsky and others construe as "s-selection" features (i.e., SEMs) might in fact turn out to be "c-selection" features (or again what I have been calling SYNs) that have certain indirect yet predictable semantic consequences.

2.4.2. Projectionism vs. Constructivism

Nothing terribly dramatic turns on this last conjecture, but it squares more evenly with a recent challenge to Chomsky's Projectionism known as *Constructivism*, variants of which can be found in Halle & Marantz (1993), Hale & Keyser (1993, 2002), Harley & Noyer (1999), Borer (2005a/b), and Ramchand (2008), and which at moments Chomsky himself appears sympathetic toward. Constructivists point out that the extreme category flexibility exhibited by most open-class roots with respect to their possible insertion frames indicates that they are inherently category-*neutral*.[54] Or at any rate most open-class Words can be "coerced" by the syntax into playing non-canonical grammatical roles by a derivational process known as *conversion*. Observe, for example, how easily/naturally the italicized roots in the sentences below adapt to their syntactic environments to play the roles indicated in parentheses:

(7a) Max is *strap*ped for cash (noun → verb)

(7b) Max's *talk* was unbearable (verb → noun)

(7c) Max *google*d the book title (name → verb)

(7d) Max *brown*ed the potatoes (adjective → verb)

(7e) Max *up*ped the ante (preposition → verb)

[54] As Ramchand (2008: 10) puts it, under these views "The root is the only lexical category."

(7f) Max has a *run*ny nose (verb → adjective)

(7g) Max can *conceiv[e]*ably win (verb → adverb)

Examples like these are indeed endless.[55] The common hypothesis here is that the placement and role of open-class roots is projected not by their lexically-encoded features but rather by the functional hierarchy (or functional "sequence") into which open-class roots are inserted. In turn, the lexical category of open-class items is inherited from or otherwise determined by the functional elements that head these sequences. On more radical Constructivist accounts, open-class Words/LIs are treated as entirely featureless morphosyntactic formatives devoid of any independently specifiable linguistic meaning (*cp.*, Marantz [1997]).

Adjudicating this question of Projectionism versus Constructivism is well beyond my expertise. However, I am compelled by clear cases of category flexibility (or what might be alternatively described as *type/category polysemy*),[56] which weighs in favor of a Constructivist-style approach to morphosyntax while inveighing against Projectionism. In contrast to radical accounts, however, there is contravening evidence to suggest that at least some open-class Words as they are listed in the lexicon are not entirely meaningless, and perhaps also not entirely featureless; *cf.*, Ramchand (2008). For instance, despite widespread category flexibility some Words are stubbornly inflexible. Take the verb 'put', which is rigidly ditransitive/trivalent (i.e., it seems to require three arguments: a subject, direct object, and indirect object) as demonstrated by (8a)-(8c):[57]

(8a) Max put the book on the table

[55] For some psycholinguistic evidence to this conclusion, see Barner & Bale (2002).
[56] This phenomenon is known as "class extension" in the developmental literature; *cf.*,McCawley (1968), Clark & Clark (1979).
[57] I use asterisks here to indicate that there is no naturally recoverable interpretation for the string of words so flagged.

(8b) *Max put

(8c) *Max put the book

Such constraints are difficult to explain without supposing that relevant selectional/subcategorization criteria are encoded in the verb's lexical entry.

While these kinds of facts/generalizations need to be reckoned with, they are also clearly exceptions to the general rule that most Words are highly adaptable to the linguistic environment into which they are inserted. The null hypothesis, however, which I am inclined to accept, is that at least some Words lexically encode the kind of constraints exhibited by (5), as violations of these constraints are non-controversially syntactic in nature. Nevertheless, the crucial observation is that grammatical processes appear capable, in most instances, of either dynamically assigning category-specific features to Words or reassigning them in the course of computation as determined by their insertion frames. This outcome is not unwelcome, however, as it greatly reduces the amount of redundancy (homophony) in the lexicon that would otherwise be needed to explain facts about category flexibility. In any case, I will briefly return to this topic in Chapter 5 while just assuming as much in what follows.

To summarize the present discourse, in 1965 Chomsky considered all lexical features to be essentially syntactic in nature. Specifically, he writes (1965: 82):

> Each lexical formative will have associated with it a set of *syntactic features* (thus *boy* will have the syntactic features [+Common], [+Human], etc.). [Chomsky's emphasis].

However, by 1986 Chomsky had begun making a distinction between c-selection and s-selection features. And, indeed, Chomsky (1995) considers the possibility that all c-selection/SYN features are predictable from and thus reducible to lexically-encoded s-selection/SEM features. Yet considerations raised by Grimshaw (1979) and Pesetsky

(1982) cast doubt on a full reduction. In particular, Chomsky concludes that idiosyncratic Case features, which correlate with the classical, positionally-defined grammatical relations "subject of" and "object of," are not entirely eliminable; Chomsky (1995: 32) observes:

> …it's important to note that formal syntactic specifications in lexical entries have not been entirely eliminated in favor of semantic ones. Whether or not verbs assign objective Case is, as far as is known at present, a purely formal property not deducible from semantics. While much of c-selection follows from s-selection, there is a syntactic residue, statable, if Psestsky is correct, in terms of lexically idiosyncratic Case properties.

Again eliding the details and attendant controversy, for purposes here the upshot is that the distinction between SYN and SEM features that I have been laboring to motivate appears to be warranted.

More specifically, the current proposal marks a clear theoretical distinction between features that have purely syntactic consequences versus those that have direct semantic significance. For while the former may have semantic "reflexes," by hypothesis they are not themselves directly interpretable by C-I. By contrast, while "bona-fide" SEM features such as [±Abstract], [±Human], and [±Animate] may have certain syntactic reflexes, they are directly interpretable at the interface to C-I. This proposal is attractive because it helps distinguish intuitions about syntactic anomalies such as (4) above (*Max declared his innocence the deed) from semantic anomalies such as that engendered by (6) above ("Sincerity may frighten justice"). For again on my view the anomaly in (6) owes not to the violation of a grammatical rule but rather to conceptual difficulties in constructing a coherent thought from the linguistically generated meaning assigned to that particular string of Words.

2.5. On the nature of lexical meanings

Setting aside facts that are exclusively syntactic in nature, the main concern of this chapter is again to characterize the nature of lexical meanings. To this end, and with minor elaborations, I shall adopt Chomsky's conception of lexical meanings/SEMs as intrinsic properties of Words that are understood, collectively, by interpreters as *instructions* to activate semantically related *extralinguistic concepts*. In turn, we can think of the meaning of a sentence as an instruction for how to construct a thought which, when all goes well, results in activation of roughly the same thought that its speaker intended to communicate. One of course wants to know more precisely what these meaning-instructions consist in. I will offer a more detailed analysis of the notion of "meanings-as-linguistic-instructions" in Chapter 5. But by way of introduction, I find it helpful to frame the idea in terms of the relation between Words (or Word-sounds) and the concepts they *lexicalize*.

2.5.1. On concept lexicalization

First, I take for granted that competent speakers typically use/utter individual words of natural language to express basic lexical concepts (i.e., roughly "word-sized" concepts). The English word 'dog', for example, is typically used to express the concept DOG(x). Conversely, a listener's interpretation of 'dog'-utterances reliably causes him or her to think about dogs, presumably in consequence having activated his/her DOG(x) concept. Such commonsense observations invite the empirical hypothesis that the *meaning* of the word 'dog' in the idiolects of most competent English speakers is somehow semantically related to their concept DOG(x). Indeed, a widely held view among cognitive and developmental psychologists is that the meaning of a word *just is* (i.e., is

constituted by) the concept that it *lexicalizes*, where concept lexicalization is said to establish a fixed association in long-term lexical memory between words and whatever extralinguistic concepts speakers associate with their meanings.[58, 59]

For example, typical learners of English typically lexicalize their concept DOG(x) under the Word 'dog' because they are typically attending to dog-thoughts (with their DOG(x) concept) at or around the time of acquisition. By noticing, perhaps only tacitly at first, the systematic correlation between 'dog'-utterances and dog-thoughts, learners eventually infer that the Word 'dog' is used/uttered by competent English speakers to express *their* DOG(x) concept. Put differently, we might say that competent speakers use the Word 'dog' to designate dogs by means of expressing their concept DOG(x). And this is evidently what learners eventually figure out. Or as Paul Bloom (2000: 8) writes: "the child is looking at a dog, someone says "dog," and she somehow connects the word with the object." At some point, so the story goes, neurology eventually takes over and the learner's DOG(x) concept becomes lexicalized (memorized, roughly speaking) under (or with) the Word 'dog'. Once lexicalized, learners then begin to utter 'dog' for themselves

[58] Paul Bloom (2000: 89) writes, for example, that "Learning a word involves mapping a form, such as the sound "dog," onto a meaning or concept, such as the concept of dogs." Murphy (2002: 385) writes "By word meaning, I mean quite generally the aspect of words that gives them significance and relates them to the world… words gain their significance by being connected to concepts." Jackendoff (1999: 306) characterizes language as "a vehicle for expressing concepts." Margolis & Laurence (1999:4) remark that "it's common to think that words in natural languages inherit their meanings from the concepts they are used to express." Snedeker (2009: 503) claims that "All theorists agree that children must learn the mapping between the phonological form and the concept." Pinker (1994: 85) says that "each person's brain contains a lexicon and the concepts they stand for…" Relatedly, Fodor (2008: 198) states that "what an English sentence means is determined, pretty much exhaustively, by the content of the thought it is used to express."

[59] As Chomsky (2000a: 61) himself puts it, one of the child's first tasks in acquiring her native language is the assignment of "labels to concepts," whereby "labels" I take him to mean a word's phonological form. However, Chomsky says little about the nature of concepts. Indeed, as Fodor (2008: 102, fn.2) observes: "It's a question of some exegetical interest why Chomsky so regularly avoids this issue. I suspect it's because he takes some sort of 'theory theory' of concepts more or less for granted." However, Chomsky (2003: 279) explicitly denies this charge.

to express their concept DOG(x).[60] In the reverse direction, subsequent 'dog'-utterances will, *ceteris paribus*, eventuate in the activation of its interpreter's DOG(x) concept such that when one hears the Word-sound \'dȯg\ one naturally thinks DOG(x).

Notice that while Chomsky describes Words as arbitrary pairings of Word-sounds with their meanings, it is widely agreed that Word acquisition involves more than mere associative learning (although associative learning may be a necessary stage of Word learning). For instance, learning that 'dog' is used to designate dogs (or to express DOG(x)) is not merely recognizing and then memorizing that dogs and 'dog'-utterances frequently co-occur. For if this were the case, learners would often fail to get the Word-to-world mapping right—as Quine observed, there are simply too many contingencies to sort out, especially in highly ambiguous contexts and in cases of displaced reference. Rather, young children quickly become aware of the fact that Words are devices of symbolic reference that can be used to refer to things even when those things are not perceptually present/salient. In addition, the learning task often requires joint attention between speaker and hearer in order to triangulate the intended object of reference. Here, even pre-linguistic infants appear to be fantastically adept at using extralinguistic cues such as eye gaze and manual gestures to disambiguate what caregivers are using unfamiliar Word-sounds to "stand for." Moreover, for such cues to be effective Word-learners must be aware of the speaker's referential intentions—i.e., they must represent speakers as having the intention to deploy these noises as symbols of the items being jointly attended to. In other words, a very young age language learners appear to

[60] I'll say more about the nature of basic/lexical concepts in Chapter 3. But for now notice that I am taking there to be just one concept DOG(x) that is shared by everyone who acquires it in virtue of each its tokens having a common content in Fodor's (1998) sense of a semantically primitive mental representation being "locked to" its referent.

recognize that the uttered symbol is produced with the intention of representing something other than the symbol itself.[61]

I will elaborate on these last few points in Chapter 4. But for now such observations are again suggestive of the fact that lexical meanings somehow relate Words to their referents by way of the concepts those Words lexicalize and are subsequently used to express. As mentioned, an even stronger claim is that the meaning of a Word *just is* the concept that it lexicalizes (or perhaps rather its representational content). More specifically, the stronger claim is that the Word 'dog' designates dogs because (i) the Word 'dog' lexicalizes the concept DOG(x), and (ii) the concept DOG(x) represents dogs. On this view, in other words, the semantic relationship between Words and the concepts they lexicalize/express is strictly *one-to-one*. Now, while one can agree that lexical meanings somehow relate Words to the concepts they lexicalize, there are several reasons to resist this stronger claim, a few of which I mentioned in Chapter 1. One very good reason, however, owes to the problem of category flexibility described above. For if the concepts expressed by Words depend on their grammatical role, and grammatical roles are context-sensitive, then so is the potential range of concepts that a given Word can be used to express. If, by contrast, lexical meanings are rigidly anchored one-to-one to particular concepts, it becomes quite difficult to explain the phenomenon of category flexibility—or, that is, again without assuming massive redundancy/homophony in the lexicon. There is, however, a second related obstacle.

[61] Or as Bloom (2000: 55) puts it "When children learn the meaning of a word, they are—whether they know it or not—learning something about the thoughts of other people.

2.5.2. The problem of lexical polysemy

In addition to category flexibility, there remains the problem of what I will call *intra-category lexical polysemy*.[62] To borrow Chomsky's favorite example, the English Word 'book' can be used as a common noun to designate either a physical copy of Tolstoy's *War and Peace* or its abstract content, and in some contexts perhaps both simultaneously. Chomsky remarks:

> One might ask whether these properties are part of the meaning of the word "book" or of the concept associated with the word... Either way, some features of the lexical item "book" that are internal to it determine modes of interpretation of the kind just mentioned.

In addition to its category label, the kinds of properties that are thought to determine a noun's "mode of interpretation" are s-selection/SEM features, including those discussed earlier such as [±Count], [±Abstract], [±Animate], [±Human], and [±Artifact], which are again classified as SEMs because by hypothesis they are directly interpretable at the interface to C-I.

For example, the Word 'book' as used to designate a physical object will include the feature-markings [+N, +Count, -Abstract, -Animate, +Artifact]. So marked, the meaning of 'book' in this context will be understood by interpreters as an instruction to activate a concept of *countable, concrete, inanimate, humanly-made objects*. But notice that there are lots of concepts, and corresponding concept-Words, that satisfy these criteria yet which have nothing to do with books. The Word 'teapot', for example, expresses a concept of countable, concrete, inanimate, artifactual objects. Yet 'teapot' clearly cannot be used in ordinary discourse to talk about books, nor *vice versa*. Thus, we

[62] I call these phenomena "problems" because arguably any theory of lexical semantics must accommodate them.

need of a conception of lexical meaning that not only accounts for what Words mean but also what they *cannot* mean. The general idea, as suggested earlier, is that each use of a Word involves the selection of features from the Numeration, some of which are fixed by the Word's lexical entry and others are derived as dictated by the linguistic context. Thus, the exact feature configuration for a given use of a Word may vary from context to context, or even from utterance to utterance.

While Chomsky is rarely explicit about this, he seems committed to the idea that the lexical entries for inherently polysemous Words like 'book' encode a stable semantic feature that constrains their possible interpretation *in all contexts* to all and only those concepts that are coherently expressible with that Word. For example, this feature would rule out the use of 'book' to express, say, the concepts CUMQUAT(x) and TURNSTILE(x) in ordinary conversation. Below I offer a formal Tarskian analysis that captures such constraints on the context-invariant aspect of a Word's meaning in terms of their satisfaction conditions. But in purely psychological terms, a plausible hypothesis, as briefly introduced in Chapter 1, is that such constraints are encoded in the lexical entries for Words as a pointer to an address in extralinguistic conceptual memory around which a potential *family* of formally distinct yet analytically related concepts all cluster; e.g., BOOK(x)$_{CONCRETE}$ and BOOK(x)$_{ABSTRACT}$.[63] By hypothesis, these pointers are established as

[63] The notion of a "family" of analytically related concepts borrows partly from Paul Pietroski (p.c.), I believe by way of Chomsky (via Wittgenstein). However, a similar notion also appears frequently in the work of contemporary native German-speaking linguists and philosophers of language, which falls under the general heading of "two-level semantics" (not to be confused with the so-called "two-dimensional semantics" of Jackson, Chalmers, et. al.), and particularly as addressed to the phenomenon of lexical polysemy. *Cf.*, e.g., Bierwisch (1983), Ruhl (1989), Bluntner (1998), Dolling (1995), and Konerding (1997). Unfortunately (for me), much of this work is either published in German or not readily accessible through American library systems. And so the references cited here derive mainly from reports in Pethõ (2001).

part of the lexicalization process and thereby fix the semantic relationship between Words and the concept(s) they lexicalize.

As a way of distinguishing these pointers from other posited SEM features, let me introduce the label 'CON' to indicate one aspect of a Word's context-invariant meaning that circumscribes the range of *CON*cepts that it can coherently be used/uttered to express in ordinary discourse. Now, it may be that CON points directly to some "core sense" of a given Word-form from which each of its subordinate senses derive. This core sense may, for instance, be the first concept lexicalized, or perhaps the most frequently activated. But either way, the idea is that activation of this core sense triggers a spreading chain of activation to its other subordinate senses via a network of cognitive associations (*cf*. Klepousniotou, et.al., 2008). Alternatively, Pylkkänen and colleagues (2006: 97) argue in favor of a so-called "single entry view" according to which "Related senses connect to same abstract lexical representation, but are distinctly listed within that representation."[64] In other words, on this latter view each sense of a Word is said to be listed under a single lexical entry. I will return to some of these details in Chapter 5. But however lexical polysemy is represented, several recent psycholinguistic studies confirm long-standing evidence that many (perhaps all) senses of a Word are typically co-activated in the course of its interpretation.[65]

Let me pause to mention that most parties to this debate presume that lexical polysemy is a property of *Words* and that the various senses of a polysemous Word are

[64] Similar findings are reported in Rodd, Gaskell, & Marslen-Wilson (2002), Beretta, Fiorentino, & Poeppel (2005) and Rubio-Fernandez (2008), among others.

[65] Gorfein (2002: 5) writes, for example, that "More than 30 years of research on lexical access has led to the conclusion that for some substantial portion of the time, multiple meanings of a word are active after the word is processed. A body of evidence further indicates that later in the process the contextually appropriate meaning of the word tends to be available but other meanings are no longer active."

represented as such in the lexicon. As I argue in in Chapter 5, however, a more empirically plausible hypothesis is that lexical polysemy is in fact an extralinguistic conceptual phenomenon adducible to the analytic relations between the concepts that polysemous Word-sounds lexicalize. Assuming the latter, one can hypothesize that Words themselves are *monosemous*—that their CON feature points to a single address in extralinguistic conceptual memory around which a potential family of concepts are interconnected through a network of associative links. In this way, we can characterize lexical meanings as linguistically-specified instructions to activate *one or more of potentially several* analytically related extralinguistic concepts (or what we might think of as related *senses* or *polysemes* of that Word).[66]

As also suggested in Chapter 1, precisely which concept/sense gets "selected" as appropriate to a given context doubtlessly depends on a number of factors, both linguistic and extralinguistic. But in most cases I assume that the meaning of a Word (i.e., its SEM features, including what I have labeled 'CON') in combination with the linguistic context in which that Word appears will be sufficient to identify which concept its utterer intended to express.[67] With respect to present concerns, however, the relevant consequence is that the meaning of an inherently polysemous Word such as 'book' cannot be identified with any one concept that it lexicalizes.[68]

From a philosophical perspective, one idea of the late Wittgenstein that Chomsky seems to embrace is that the meaning of an inherently polysemous Word such as 'book'

[66] For all one knows, Chomsky borrowed this idea from Strawson (1950) who held a similar view about referential terms.
[67] And when that's not enough, speakers can utilize metalinguistic beliefs about what a given Word can be used to express, as I argue in greater detail in Chapter 5.
[68] A related problem, which I call the problem of conceptually underspecified Words, is taken up in Chapter 3.

makes available multiple senses of that Word that exhibit a kind of "family resemblance." To endorse this much is not however to adopt Wittgenstein's skepticism about the attribution of rule-following in a private language as a way of explaining a speaker's linguistic competence. For again according to Chomsky, we are to think of the meanings of Words, and of linguistic expressions more generally, as being, or at least being understood as "rules" or "instructions for thought and action."[69] Yet commonsense dictates that the same rules/instructions can often be satisfied, or obeyed, or executed in a number of different ways. In fact, if I understand Chomsky correctly that's just what it is for a Word to be polysemous. For in the present example there are at least two perfectly good ways (actually more than two)[70] of executing the meaning-instruction encoded by the lexical entry for 'book', which is by activating either BOOK(x)$_{CONC}$ or BOOK(x)$_{ABS}$.

2.5.3. On the nature of Words and their lexical meanings—Summing up

In summary of the discussion to this point, I have been advocating a technical conception of Words that are internally represented by (or within) I-languages as particular instantiations of the complex mental structure <PHON, SYN, SEM>. In a modest departure from standard Chomskyan terminology, though keeping to the spirit of his view, I have introduced the label 'CON' as way of talking about the context-invariant aspect of a Word's meaning that restricts its range of application to all and only those concepts that are coherently expressible with that Word. To be clear, however, I am assuming that CON is one of perhaps several other SEM features encoded by a Word's lexical entry and that taken together constitutes its linguistic meaning. In terms of processing, it is the CON feature of a Word that points interpreters to the right conceptual

[69] Chomsky (2000a: 9).
[70] One can also book a ticket, a wager, an appointment, and so on.

"neighborhood." From there, other SEM features conspire with contextual factors to narrow down the search. And when all goes well, the search is constrained to precisely the concept that its speaker intended to express.

As cited earlier, Chomsky maintains that "LIs may be decomposed and reconstructed in the course of computation" and, moreover, that "any feature change yields a different LI." In my terminology, this means that any difference in *lexically-encoded* features yields a different *Word*. And in terms of lexical meanings, any difference in lexically-encoded SEM features yields a different meaning. Thus, in response to our opening question which asked under what conditions two utterances are tokens of the same Word, my proposal is this: Whenever those utterances are registered by a speaker's I-language as being type-identical with respect to each of their lexically encoded PHON, SYN, and SEM features.

2.6. On referential uses of Words

I have to this point been assuming a distinction between the linguistically-encoded meanings of Words and their denotational or referential uses in discourse. In this section I will sketch a hypothesis about such uses, focusing in particular on referential uses of proper names. For as discussed in Chapter 1, names play a central role in my core thesis about metalinguistic-semantic competence (MSC) and its relation to semantic theory. And the semantic theory for names developed below will be assumed in later chapters.

2.6.1. Lexical expressions (LEs)

To repeat, I am assuming that Words are best characterized (with noted exceptions) as lexical roots that appear as terminal nodes in syntactic trees. As observed earlier, open-class Words/roots combine with other LIs to create morphologically complex

expressions. In this section I argue that some Words—and nouns in particular (both common and proper)—also combine with syntactic indices to generate non-terminal *lexical expressions*, or *LEs* for short. In a sense to be made more precise in Chapters 6 and 7, this formal distinction has implications for our ordinary conceptions of Word-meanings.

2.6.2. Referential indices

It has become commonplace nowadays among semanticists to posit various kinds of covert (unpronounced) syntactic elements—including variables and/or indices—in the logical forms of linguistic expressions to account for grammatically controlled semantic phenomena such as anaphora, ellipsis, theta-linking, quantifier scope and domain restriction, among others.[71] Most relevant to my purposes here, Baker (2003) argues, for example, that *referential indices* attached to common nouns are what distinguish nouns from other lexical categories. I refer readers to the source for the full technical details. But in brief outline, Baker's argument derives from Geach (1962) and Gupta (1980) according to whom "only common nouns have a component of meaning that makes it legitimate to ask whether some X is the same (whatever) as Y." And this is to suggest that only nouns have *criteria of identity*—they "provide standards of sameness by which we can judge whether X is the same as Y."[72]

[71] Other applications include the contextual provisioning of comparison classes for comparative adjectives, and likewise for the contextual precisification of gradable adjectives.

[72] Baker acknowledges Chomsky's recent injunction against syntactic indices as superfluous/redundant to computation.[72] However, Baker claims that his reasons for invoking indices are not those to which Chomsky objects. Or in any case, he writes: "It does not actually matter to my theory whether these indices are present throughout the computation of a linguistic structure. A legitimate alternative would be that these indices are added at the conceptual-intentional interface, just beyond LF. The substance of my theory can thus be made consistent with Chomsky's (1995) view that indices are not part of the linguistic representation proper. I nevertheless include indices freely in my syntactic representations, because I do not know any compelling reason to say they are not there and because it makes the representations more

As Baker conceives them, nominal indices consist in an ordered pair of integers yielding complex expressions of the form '$X_{\{j,\ k\}}$' which correspond to the interpretation "j is the same X as k." More specifically, nominal expressions of the form '$X_{\{j,\ k\}}$' express a two-place equivalence relation between the semantic value (or referent) of an indexed noun and other co-occurring, coindexed expressions.[73] The relevant idea is that the distribution of nominal indices is regulated by the grammar which thereby determines formal constraints on the coreference (or non-coreference) of their tokens. In short, "anything that has a referential index must be coindexed with something else in the structure," where coindexation implies coreference.[74]

A similar but differently motivated view is defended by Fiengo & May (1998, 2006) to account for the referential codependency between type-identical tokens of proper names. I return in Chapter 6 to consider their proposal in greater detail. But the point for now is that Fiengo & May cite convincing reasons for distinguishing *names* from the *syntactic expressions* of those names in a discourse. According to them, names are lexical items (or roughly speaking, *Words* in my terminology) that "occur in" or are "contained by" non-terminal expressions of those names. Importantly, Fiengo & May contend that names themselves are semantically empty—they have no independently specifiable meaning/reference outside of a discourse.

More specifically, Fiengo & May treat names merely as syntactic formatives that combine with referential indices to create complex nominal expressions of the form '[$_{NP}$

---

explicit. I leave the exact status of these indices at the different stages of linguistic computation open for further conceptual reflection and empirical research."

[73] For the record, however, Baker qualifies that "readers are invited mentally to reduce my ordered pairs to a simple integer if they like. But given that the bearing of a referential index is underwritten by the lexical semantic property of having a criterion of identity in my view, and since identity is inherently a two-place relation, I assume that pair-indices ultimately make more sense conceptually.

[74] Baker's account is especially well-suited to explain robust facts about grammatically-determined referential dependencies between anaphors and their nominal antecedents.

Aristotle]$_i$', where the index '*i*' tracks the syntactic identity (or difference) of its tokens (again relative to a context/discourse). Since type-identical tokens are necessarily covalued, i.e., as a function of the grammar, coindexation of nominal expressions determine their coreference. For example, the two coindexed expressions of the name 'Aristotle' in (9) below indicate that they are type-identical and thus covalued (or coreferential):

(9) [$_{NP}$ Aristotle]$_1$ is [$_{NP}$ Aristotle]$_1$

Indeed, it is a competent speaker's awareness of *expression* identity that renders trivial the *referential* identity asserted by (9). By contrast, *non*-coindexed expressions may or may not corefer as determined by speaker-intentions. For instance, the tokens of '[$_{NP}$ Aristotle]$_i$' in (10)-(12) are type-distinct, as determined by their non-coindexation, and are therefore free to refer to different individuals:

(10) [$_{NP}$ Aristotle]$_1$ was a Greek philosopher

(11) [$_{NP}$ Aristotle]$_2$ was a Greek shipping magnate

(12) [$_{NP}$ Aristotle]$_1$ is not [$_{NP}$ Aristotle]$_2$

The NPs in (9)-(12) are roughly equivalent to what above I called lexical expressions, or again *LEs* for short. My analysis of LEs is similar to Fiengo & May's yet avoids some dubious consequences of their view. In particular, I argue that names themselves have fixed (i.e., context-invariant) meanings that restrict their range of application to only things so-called. In this way, we can think of the meanings of names (i.e., their CON features) as imposing restrictions on the domains over which the LEs that contain them range. Furthermore, I argue that the attachment of referential indices is *optional* as determined by speaker-intentions (but without venturing a guess as to precisely how speaker-intentions determine the outcome, one way or the other).

Following Fiengo & May, it will be a strategy of this work to specify the semantics of LEs in terms of Tarskian satisfaction conditions. With respect to notation, rather than using simple numerical indices I will employ syntactically complex variables (x_1, x_2, x_3… x_n) to render their status as independent syntactic constituents more perspicuous. The full details appear in Appendix A at the end of this Chapter. But as a sample, one can stipulate that the bare NPs '[$_{NP}$ [$_{PN}$ Cicero]]' and '[$_{NP}$ [$_{PN}$ Tully]]' combine with a referential index, [x_n], $n > 0$, to form are well-formed, non-coindexed yet coreferential LEs whose Tarskian assignments relative to a sequence, σ, are given by (13) and (14):[75]

(13a) $g(x_i)^{\sigma} = \sigma(i)$, $i > 0$

(13b) $g([[_{NP}\ [_{PN}$ Cicero]][x_i]])$^{\sigma}$ = λx: x satisfies 'Cicero' iff $x = \sigma(i)$ & $\sigma(i)$ is called 'Cicero'

(14a) $g(x_k)^{\sigma} = \sigma(k)$, $k > 0$

(14b) $g([[_{NP}\ [_{PN}$ Tully]][x_k]])$^{\sigma}$ = λx: x satisfies 'Tully' iff $x = \sigma(k)$ & $\sigma(k)$ is called 'Tully'

According to (13b), the semantic value of the LE '[[$_{NP}$ [$_{PN}$ Cicero]][x_i]]' relative to any sequence, σ, is the i-th object in σ, and (14b) says that the semantic value of '[[$_{NP}$ [$_{PN}$ Tully]][x_k]]' is its k-th object. Since 'Cicero' and 'Tully' are typically used in reference to the same individual, it will turn out in typical cases that $\sigma(x_i) = \sigma(x_k)$ = Marcus Tullius Cicero, for any sentence in which those names appear are true.

To repeat a point made in Chapter 1, the category label 'PN' attached to names ensures that their tokens apply only to nameable things, whereas the lexically-encoded meanings of particular names further restricts their satisfaction conditions to all and only

[75] Evidence suggests that names in English always combine with a covert determiner such as 'the' or 'that', in which case it would be more appropriate to label these phrases as DPs rather than NPs. However, I needn't take a stance on the matter here.

those objects in the domain that are in fact so-called. The indices with which names optionally combine ensure that the LEs containing them refer uniquely to a single individual relative to a context/discourse (assuming that there is a referent that answers to the expression).[76]

I again refer readers to Appendix A below for the full technical details. But in philosophical terms, as also briefly mentioned in Chapter 1, I am endorsing a view of proper names similar to Burge (1973) according to which names behave, semantically, like ordinary one-place predicates. As Burge argues, names function like demonstratives—expressions that have fixed meanings but variable reference. In short, while it appears that names have meanings that range over objects/individuals so-called, I nonetheless agree with Fiengo & May that names in isolation do not themselves *refer*. And so in this sense proper names in natural languages are more like what Kaplan (1990) calls "generic names." Rather, on the view that I am endorsing it is the syntactic *expression* of a name relative to a context/discourse that facilitates its use as a device of direct, singular reference. However, names can also be used *predicatively*, which is to say without an index, in which case they are understood as expressing their lexically-encoded meanings, which is to say the property of being called by the name 'PN'.

One benefit of this proposal is that it stands to explain the difference between referential versus predicative/attributive uses of names. For instance, if one intends to refer to Aristotle Onassis, one might employ the indexed LE '[[$_{\text{NP}}$ [$_{\text{PN}}$ Aristotle]][x_i]]'. However, if one intends to use a name with no particular individual in mind, one would instead use a bare (i.e., *non*-indexed) LE. For instance, by hypothesis utterances of the

[76] There may be exceptions to this rule, as in Donnellan-style cases where a proper name is used successfully to refer to an individual not so-named.

question-sentence "Which Aristotle are you talking about?" contains as a constituent the non-indexed LE '[$_{NP}$ [$_{PN}$ Aristotle]]'. Similarly, one might introduce a nonce name into the discourse by saying "Take your average John Doe" whereby one employs the non-indexed LE '[$_{NP}$ [$_{PN}$ John Doe]]' as a way of talking about no one in particular. Or to borrow a related example from Kent Bach (2002: 87):

> Suppose you and your friend discover a briefcase containing a large amount of money and quickly put the money in your shopping bag. You close the briefcase, put it down, and notice the name 'Cassius King' on the nameplate. You say to your friend, "Cassius King won't be happy, but at least he'll have his briefcase."

Here, the speaker is not using the name 'Cassius King' referentially but rather attributively, which is to say *sans* index. The same remarks apply equally to attributive/predicative uses of definite descriptions (e.g., '[$_{DP}$ [$_{DD}$ the average taxpayer]]').[77]

As importantly, and as also outlined in Chapter 1, this analysis of names can help to explain the informativeness of true identity statements of the form 'α is β', where 'α' and 'β' are coreferential proper names. Illustrating this point is part of the project of Chapter 7. But as a second substantive assumption of this work, I devote Chapter 3 to a review of Jerry Fodor's conception of concepts and their representational contents. Several features of this discussion apply to my overall thesis. But perhaps most saliently I will be using Fodor's view to ground my conception of the nature of Word concepts, or again *concepts*$_W$ as I am distinguishing them. In general, a Fodorian theory of concepts coupled

[77] More generally, I suspect that a speaker's prerogative to deploy LEs with or without a referential index stands to explain the tendency of ordinary speakers to conflate the linguistic-semantic properties of *Words* with the referential properties of their *lexical expressions* in a discourse.

with a Chomskyan conception of Words I believe improves the overall plausibility of my conception of Word-concepts and their role within a theory of semantic competence.

Appendix A

1. Encoding referential indices in a Tarskian-style semantics for natural language

Recall that in Tarski's system open sentences (or sentential functions) are said to be satisfied by *sequences* of ordered *n*-tuples of objects in the domain of discourse. More specifically, we can say that each free variable of an object language *L* stands in one-to-one (surjective) correspondence to positions in a sequence of the general form σ = <1, 2, 3, …, *n*>, where the integers here serve as placeholders for corresponding objects in that sequence. This correspondence can be defined by an assignment (interpretation) function, *g*, that recursively maps sequence-objects onto corresponding variables that occur in open sentences. Accordingly, the system must allow for as many formally distinct variables as there are object positions in the sequence, and possible sequence-variants, which as Tarksi allowed may be infinitely many.

To capture these distinctions more precisely, let me introduce the following axioms/definitions:

- Let x be a variable of *L* that ranges over sequence-objects.
- Let (x ^ i) describe a recursive operation on variables capable of generating infinitely many syntactically complex variables of the form $<x_1, x_2, x_3, \ldots, x_i>$, such that for any sequence, σ, x_1 corresponds to the first object in σ, x_2 corresponds to the second object, x_3 to the third, and so such that x_i corresponds to to the i-th object for any $i \geq 1$.
- Let Words of natural language, designated by α, $=_{df}$ a class of naturally acquirable, syntactically primitive linguistic expressions whose tokens are

particular instantiations of the generic mental structure {PHON, SYN, SEM} as characterized in Chapter 2 above.

- Let us further stipulate that Words combine, syntactically, with variables (i.e., referential indices) via the expression-forming operator '^' to generate infinitely many lexical expressions (LEs) of the form '$[_{XP}\ [_{XL}\ \alpha][x_i]]$', where 'XP' is a phrasal label (e.g., NP, VP, AP, etc.), 'XL' is a lexical category label for the Word 'α', and the square brackets reflect the expression's constituent structure.[78]
- ***NB*****:** *Pace* Fiengo & May I distinguish Words from Tarskian variables by stipulating that the former *apply to* a restricted range of sequence-objects as determined by their category labels, whereas the latter *range over* all sequence-objects, though as restricted by lexical constituents with which they combine (see below).
- Thus, we can let the following formula be an axiom of the system,

  $g(\text{'}[_{XP}\ [_{XL}\ \alpha][x_i]]\text{'})^{\sigma} = \sigma(x_i)$, for all $i \geq 1$

  where g is function that assigns semantic values to variables, where '$[_{XL}\ \alpha]$' is a labeled Word that applies to a restricted range of sequence-objects, where 'x_i' is again a syntactically complex variable that ranges over sequence-objects so restricted, and where 'XP' is a phrasal category generated by the syntax as the result of inserting the term '$[_{XL}\ \alpha]$' into a non-terminal node of the emerging structure.

As briefly mentioned in Chapter 1, and as will be important in later chapters, let me explain the interpretation of this notation using proper names, where proper names (or proper nouns) are understood to constitute a particular class of *Words* in the technical sense characterized above.

[78] As suggested earlier, Words might first combine with other Words to form syntactic compounds, incorporations, and portmanteaus on the assumption that such constructions are treated by the grammar as individual Words.

First, let us assume for example that the formulae '[$_{PN}$ Cicero] ^ [x_1]' and '[$_{PN}$ Tully] ^ [x_2]' generate well-formed LEs whose Tarskian assignments relative to a sequence σ are given by (1) and (2), respectively:[79]

(1) $g(\text{'}[_{NP}\ [_{PN}\ \text{Cicero}][x_1]]\text{'})^{\sigma} = \sigma(x_1)$

(2) $g(\text{'}[_{NP}\ [_{PN}\ \text{Tully}][x_2]]\text{'})^{\sigma} = \sigma(x_2)$

According to (1), the semantic assignment of '[$_{NP}$ [$_{PN}$ Cicero][x_1]]' relative to any sequence, σ, is the first object in σ, and (2) says that the semantic value of '[$_{NP}$ [$_{PN}$ Tully][x_2]]' is its second object. Stated in terms of satisfaction conditions, we can conditionalize (1) and (2) as follows:

(1') $g([[_{NP}\ [_{PN}\ \text{Cicero}]][x_i]])^{\sigma} = \lambda x$: x satisfies 'Cicero' iff $x = \sigma(i)$ & $\sigma(i)$ is called 'Cicero''[$_{NP}$ [$_{PN}$ Tully][x_2]]'$^{\sigma}$ is satisfied by $\sigma(x_2)$ iff $\sigma(x_2)$ is called 'Tully'

(2') $g([[_{NP}\ [_{PN}\ \text{Tully}]][x_i]])^{\sigma} = \lambda x$: x satisfies 'Tully' iff $x = \sigma(i)$ & $\sigma(i)$ is called 'Tully'

Once again, I assume that the lexical category label 'PN' (for proper name) ensures that name-tokens apply only to things so-called (again without saying what it is *to be* so-called). In other words, the label 'PN' formally restricts the satisfaction conditions of the nominal LEs in (1') and (2') above to all and only those objects in the sequence that are so-called.[80]

[79] Evidence further suggests that names in English always combine with a covert determiner such as 'the' or 'that', in which case it would be more appropriate to label these phrases as DPs rather than NPs. However, I won't take a stance on the matter, and need not for purposes here.

[80] If we assume a Kaplanian extension to account for "pure" indexicals such as 'I', 'here', and 'now', we might reserve a small fragment of the Tarski sequence for speaker, place, and time. This fragment is roughly what Fiengo & May (2006) refer to as a "context." Specifically, they write (p.27) that "contexts are sequences of (at least) persons, places, and times…" When speaking of sequences, then, the presence of this fragment shall be implied.

As also mentioned in Chapter 1, for purposes of describing a speaker's semantic psychology I will specify the sequence-independent (i.e., context-*invariant*) meanings names such as 'Cicero' and 'Tully' along the following lines:

(3) CON(√cicero) = *activate@cicero* → {CICERO(w), IS-CALLED(*C*), IS-CALLED(*C*, *C*), CICERO}

(4) CON(√tully) = *activate@tully* → {TULLY(w), IS-CALLED(*C*), IS-CALLED(*C*, *C*), TULLY}

With respect to the concepts$_W$ IS-CALLED(*C*, *C*) and IS-CALLED(*C*), the restricted higher-order variable *C* can only be saturated by metalinguistic concepts$_W$ such as CICERO(w), whereas *C* in IS-CALLED(*C*, *C*) must be saturated by singular concepts of objects/individuals such as CICERO, TULLY, and so on. Alternatively, one could specify the following sequence-dependent *application conditions* for the name '[$_{PN}$ Cicero]' and '[$_{PN}$ Tully]' as follows:[81]

(3') '[$_{PN}$ Cicero]'$^{\sigma}$ applies to $\sigma(x_1)$ iff $\sigma(x_1)$ is called 'Cicero'

(4') '[$_{PN}$ Tully]'$^{\sigma}$ applies to $\sigma(x_2)$ iff $\sigma(x_2)$ is called 'Tully'

Similarly, (5) and (6) below are ways of specifying assignments of values to variables for two formally distinct LEs that contain *the same name*, with their corresponding satisfaction conditions specified by (5') and (6'):

(5) g('[$_{NP}$ [$_{PN}$ Aristotle][x_1]]')$^{\sigma}$ = $\sigma(x_1)$

(6) g('[$_{NP}$ [$_{PN}$ Aristotle][x_2]]')$^{\sigma}$ = $\sigma(x_2)$

(5') '[$_{NP}$ [$_{PN}$ Aristotle][x_1]]'$^{\sigma}$ is satisfied by $\sigma(x_1)$ iff $\sigma(x_1)$ is called 'Aristotle'

(6') '[$_{NP}$ [$_{PN}$ Aristotle][x_2]]'$^{\sigma}$ is satisfied by $\sigma(x_2)$ iff $\sigma(x_2)$ is called 'Aristotle'

[81] See Pietroski (*forthcoming*) for discussion about the distinction between application conditions and Tarskian satisfaction conditions.

2. Accommodating grammatically-determined referential relations

Other members of the class of sequence-dependent Words include pronouns and reflexives. As such, my notation is designed to accommodate the phenomena to which the binding principles of traditional GB theory were originally addressed, as characterized by the following three principles; *cf*., Chomsky (1981, 1986):

Binding Theory

(A) Reflexive pronouns are locally bound.

(B) Personal pronouns are locally free.

(C) All other NPs (including Names) are globally free.

In short, whatever theory best *explains* these facts, our notation must minimally be capable of *describing* them.[82]

For instance, the notation must make explicit the fact that the reference of 'himself' in (7) below is anaphorically dependent on its local (c-commanding) antecedent, 'Max', while 'him' in (8) is not so-bound; i.e., there is no natural reading of (8) under which the Pronoun 'him' refers to Max, and likewise the pair 'He'/'him' in (9) cannot corefer:

(7) Max helped himself

(8) Max helped him

(9) He helped him

Our system must also reflect the fact that 'his' in (10) below can be understood either as referentially bound to 'Max' *or* as taking its value deictically from some other conversationally salient male. By contrast, 'He' in (11) cannot be Max (except perhaps under forced/unnatural circumstances):

[82] Special thanks here to Paul Pietroski (p.c.) for pressing me to consider more carefully the import of these details.

(10) Max helped his mother

(11) He helped Max's mother

Similarly, the notation must be powerful/flexible enough to describe the fact that 'him' in (12) below can refer either to Max or some other contextually salient male, but not to Max, that 'himself' in (13) must refer to Max, whereas in (14) 'himself' can only refer to Max.

(12) Max wants Max to help him

(13) Max wants to help himself

(14) Max wants Max to help himself

Finally, consider (15) where the second token of 'Max' can be understood as either bound to the first token or free to be saturated by some other discourse-salient Max.

(15) Max wants to help Max

So described, we can add pronouns to the list of Words in *L* that combine with variables to form complex open formulae (LEs) such as '[$_{NP}$ [$_{PR}$ him][x_i]]', '[$_{NP}$ [$_{PR}$ she][x_i]]', and so forth, where 'PR' (pronoun) is taken to be a valid category label encoded by their lexical entries. One further assumes that certain Pronouns can combine with the reflexive formative 'self' to generate '[$_{NP}$ [$_{RP}$ himself][x_i]]', '[$_{NP}$ [$_{RP}$ herself][x_i]]', etc., where 'RP' (reflexive pronoun) is also a valid category label. With these additions we now have the resources to capture the relevant binding facts of (11)-(19), as demonstrated below, and under the assumption that the indexed variables appearing in the expressions below are part of their underlying logical forms:

(16') [$_{NP}$ [$_{PN}$ Max][x_1]] helped [$_{NP}$ [$_{RP}$ himself][x_1]]

(17') [$_{NP}$ [$_{PN}$ Max][x_1]] helped [$_{NP}$ [$_{PR}$ him][x_2]]

(18') [$_{NP}$ [$_{PR}$ He][x_1]] helped [$_{NP}$ [$_{PR}$ him][x_2]]

(19') [$_{NP}$ [$_{PN}$ Max][x_1]] helped [$_{NP}$ [$_{PR}$ his][x_1]] mother

(19'') [$_{NP}$ [$_{PN}$ Max][x_1]] helped [$_{NP}$ [$_{PR}$ his][x_2]] mother

(20') [$_{NP}$ [$_{PR}$ He][x_1]] helped [$_{NP}$ [$_{PN}$ Max][x_2]]'s mother

(21') [$_{NP}$ [$_{PN}$ Max][x_1]] wants [$_{NP}$ [$_{PN}$ Max][x_2]] to help [$_{NP}$ [$_{PR}$ him][x_1]]

(21'') [$_{NP}$ [$_{PN}$ Max][x_1]] wants [$_{NP}$ [$_{PN}$ Max][x_2]] to help [$_{NP}$ [$_{PR}$ him][x_2]]

(22') [$_{NP}$ [$_{PN}$ Max][x_1]] wants to help [$_{NP}$ [$_{RP}$ himself][x_1]])

(23') [$_{NP}$ [$_{PN}$ Max][x_1]] wants [$_{NP}$ [$_{PN}$ Max][x_2]] to help [$_{NP}$ [$_{RP}$ himself][x_2]]

(24') [$_{NP}$ [$_{PN}$ Max][x_1]] wants to help [$_{NP}$ [$_{PN}$ Max][x_1]]

(24'') [$_{NP}$ [$_{PN}$ Max][x_1]] wants to help [$_{NP}$ [$_{PN}$ Max][x_2]]

3. On the Nature of Concepts and their Representational Contents

3.1. Introduction

I have been assuming all along that humans, like our nearest non-human ancestors, possess a great many pre-linguistic concepts, some of which are doubtlessly innate and many of which are consciously accessible in human thought; though as briefly suggested in Chapter 2 some aspects of human thought is doubtlessly *non*-conceptual in nature. I further assume that a great many other concepts will be acquired throughout a human's lifetime, including his/her concepts of Words, or again *concepts*$_W$ as I am distinguishing them. While the exact nature of concepts, in general, remains highly controversial, I shall for wont of a better theory take them to be Fodorian in nature. Although here again I trust that my core thesis is largely compatible with other theories on the market. Yet as with Chomsky's conception of Words, my preference for Fodor's theory of concepts is that it renders my story about concepts$_W$ and the metalinguistic-semantic competence (MSC) they support more theoretically perspicuous and, I believe, more explanatorily powerful. In particular, Fodor's account of concept acquisition will figure into my proposal in Chapter 4 about the acquisition of lexical meanings.

For these reasons, I devote quite a bit of effort here working through the relevant details, beginning with a brief review of the central tenets of Fodor's account of basic/lexical concepts and his theory of mental representation more generally. In the latter half of this Chapter I will then demonstrate where concepts$_W$ fit into this picture, again deferring the details of how my account connects up with the acquisition of lexical meanings for the next chapter.

3.2. Fodor's bold conjecture

Theorists who endorse a representational theory of mind (RTM) also tend to agree that the best (and perhaps only) way to explain the observed productivity and systematicity of human thought is to assume, similar to natural language semantics, that its representational contents are compositionally determined by the contents of its constituent parts according to their structural (i.e., syntactic) arrangements. Under the further assumption that the basic constituents of thought are roughly "word-sized" *concepts*, Fodor's rather strict principle of compositionality holds that "the content of a thought is *entirely* determined by its structure together with the content of its constituent concepts" (Fodor [2008: 17], my emphasis). He adds (*ibid*: 19-20):

> Over the last couple of years I've become increasingly convinced that capturing the compositionality of thought is what RTM most urgently requires; not just because compositionality is at the heart of the productivity and systematicity of thought, but also because it determines the relation between thoughts and concepts. The key to the compositionality of thoughts is that they have concepts as their constituents.

While many researchers are willing to go along with Fodor's claim that the contents of complex thoughts are compositionally determined, this leaves open questions about the underlying nature of basic concepts and their representational contents.

In partial answer to this question, Fodor has insisted for decades what basic concepts *cannot* be. First, they cannot be *definitions* as maintained by the classical Aristotelian view because the vast majority of basic concepts are not definable in the sense of there being necessary and sufficient conditions on their possession and/or application. Nor are concepts *prototypes* (or *stereotypes*) because prototypes don't compose, or at least not in the way that basic human concepts do, which is to say productively and systematically. Nor can concepts be grounded in *inferential roles*,

among other reasons because the possession conditions of basic concepts hold independently of their inferential relations to other concepts.[1] One can for example have the concept DOG(x) without having the superordinate concept ANIMAL(x), and *vice versa*. Of course, while it may be *nomologically* necessary that all dogs are animals, this fact is not logically entailed by the *content* of DOG(x). Nor can such inferences be licensed by the internal constituent structure of DOG(x) because it has none. In other words, the entailment $\forall x$ [DOG(x) $\supset$ ANIMAL(x)] is not by Fodor's lights *analytic*. Thus, the content of DOG(x) is neither constituted nor determined by its inferential relation to ANIMAL(x).

Fodor's positive account of concepts decomposes into a portfolio of related sub-theses which are variously known as *conceptual atomism, informational semantics*, *referentialism,* and (*radical*) *concept nativism*. Combined, the first three of these theses constitute what Fodor calls *referential* or *informational atomism* whose truth implies the truth of (*radical*) *concept nativism.* While likely familiar to my target audience, allow me a moment to briefly remind ourselves what each of these hypotheses amounts to.

First, *conceptual atomism* is the thesis that "most lexical concepts have no internal structure" (1998: 121).[2] Specifically, Fodorian atoms are *syntactically unstructured* symbolic representations in a thinker's internal mental language (i.e., one's so-called *language of thought* or *Mentalese*). Construed as *mental symbols*, atomic (i.e., constituent-less) concepts figure into the causal-computational processes of thought in virtue of the formal properties of their tokens (which are said to be "content-preserving"

[1] These are all known as "knowledge-based accounts" of concepts; *cf*., Murphy & Medin (1985) for discussion.
[2] More precisely Fodor would say that atomic concepts have no *semantically significant* internal syntactic structure/orthography.

over inferential processes).[3] Given their unstructured nature, we can think of the purely formal properties of these uninterpreted mental symbols in terms of their physical "forms" or "shapes" (presumably as realized by neurophysiological properties of corresponding brain states). Atomic concepts, understood as mental representation-types, are thus individuated by computational processes according to the physical forms/shapes of their tokens.[4]

Construed as mental *representations*, however, concepts also have intentional properties; they are by Fodor's lights *about* things in the world.[5] In the reverse direction, worldly particulars supply conceptual atoms with their representational or semantic *contents*. More precisely, Fodor's version of informational semantics holds that conceptual contents are "constituted by some sort of nomic [counterfactual-supporting], mind–world relation" (1998: 121). These mind-world (or symbol-world) relations are in turn said to be subsumable under some as-of-yet undiscovered laws of psychology (more on this below).[6] Referential atomism adds that "The content of [an atomic] mental representation is its *referent*, and what fixes its reference is the character of its *causal* connections to the world" (2008: 216, my emphases). Atomic concepts are claimed to be semantically primitive because on Fodor's view "reference is the only primitive mind-world semantic property" (*ibid*: 16). The upshot is that Fodorian atoms are type=individuated by both their *form* and *content*. Hence, one can possess formally

[3] The inferential relations that primitive mental symbols enter into in computational processes are however not *constitutive* of their contents, which is to deny the existence of so-called "narrow content."
[4] Fodor is a token-physicalist and thus allows for multiple realization of concept-tokens over their types.
[5] Fodor (2008: 44) writes "most concepts apply to things in the world. Or, anyhow, that's what they are supposed to do." That is, Fodor appears to accept Brentano's thesis at face value.
[6] Fodor (1998: 121) adds that "Correspondingly, having a concept (concept possession) is constituted, at least in part, by being in some sort of nomic, mind–world relation."

distinct coreferential concepts that play distinct causal-computational roles in cognition without supposing that computation roles are constitutive of their content.

Lastly, *radical concept nativism* is the thesis that all atomic concepts are *innate*. This claim has of course generated more than its share of controversy over the years. Not because it implies that thinkers are born with or otherwise genetically pre-disposed to acquire specific concepts such as DOORKNOB(x) and CARBURETOR(x). Rather, to call these concepts "innate" also calls for a somewhat revisionary understanding of what innateness *is*, or what the word 'innateness' *means*. For as commonly construed, innate concepts are programmed by nature into a thinker's genetic code and thus either: (i) present at birth, or (ii) apt to be acquired some time later as a natural consequence of neurological/ontogenetic development, perhaps in coordination with certain environmental, experiential, and/or hormonal "triggers."

Now, Fodor takes as common ground among his critics that "Minds like ours start out with an innate inventory of concepts," of which he adds "there are more than none but not more than finitely many" (2008: 131). Yet whatever this stock of prenatally determined concepts happens to be, nobody in their right mind, including Fodor as I understand him, thinks that DOORKNOB(x) and CARBURETOR(x) are among them *if* innateness is understood in either of the two senses above. There is however a third, weaker sense of innateness which by Fodor's lights does apply. And this is simply to say that an innate concept is one that is *not learned*, where 'learned' is understood in the traditional Empiricist sense of involving inductive/statistical generalizations over a concept's prototypical instances. On this latter view, the output of learning is taken to include the ability to reliably sort and/or name those instances *as instances* of a particular

category.[7] However, since Fodor feels quite confident that basic concepts are not in this way learned, and provides what I take to be good arguments in support of this claim, he concludes that they must all be innate.

Notice that Fodor is not just playing games with the meaning of 'innateness'. For there is still something "genuinely" innate about the acquisition of basic concepts such as DOORKNOB(x) and CARBURETOR(x), which is a cognitive *mechanism* that has evolved in creatures like us for just this purpose; i.e., to enable minds so-constituted, under appropriate conditions and with the right kinds of experiences, to *acquire* (as opposed to learn) concepts by way of becoming metaphysically locked to their referents.[8] Precisely how this locking mechanism works in neurological terms is admittedly anyone's guess (though see below). However, Fodor is relying on an inference to the best explanation which runs something like this: Concepts are by definition intentional and *ex hypothesi* referential. In order to satisfy the explanatory demands of traditional belief-desire psychology, the referential relations between concepts and their contents must be subsumable under broad psychological laws. Fodor's wager is that these laws are embodied by an innate mechanism that facilitates concept acquisition. And in his words, since "acquiring a concept is getting nomologically [i.e., counterfactually] locked to the property that the concept expresses" (1998: 125), the mechanism that explains concept acquisition also explains why "having a concept is being locked to a property" (*ibid*.: 126).

[7] Fodor (2008: 163).

[8] To say that a biological mechanism (or mechanisms) evolved that enables minds like ours to become nomologically locked onto properties is not to say that this mechanism (these mechanisms) *evolved for that purpose*. For it is compatible with Fodor's view that this outcome was an accident of nature, or perhaps a spandrel effect in Gould's sense.

I'll return to the details of concept acquisition presently. But Fodorian considerations again arguably rule out competing theories of conceptual content grounded in definitions, prototypes, and inferential roles. Fodorian atoms do however compose to generate the *complex* (i.e., syntactically structured) mental formulae over which inferential processes are defined. To take a paradigm example, consider the complex concept BROWN-DOG(x) whose atomic constituents are presumably the monadic concepts BROWN(x) and DOG(x). Let us further assume with Fodor that the content of BROWN(x) is the semantically primitive property of being brown and that the content of DOG(x) is the property of being a dog. Fodor's principle of compositionality dictates that the content of BROWN-DOG(x) is determined by the semantic conjunction of BROWN(x) and DOG(x), which yields the intersective (i.e., complex) property of being a brown dog.[9] Or to render its structural analysis more explicit, I will say that the content of [BROWN(x) & DOG(x)] is the property *brown+dog*.

In this way, *complex* concepts can be learned, which is to say constructed/composed, so long as learners antecedently possess each of their constituents. In the reverse direction, complex concepts are also subject to decompositional analysis and may therefore bear analytic entailments. For example, Fodor would readily agree that if anything is analytically true it's that brown dogs are both brown and dogs. Likewise, it is analytic that if all dogs are animals then so are all brown dogs, and if no dogs are marsupials then no brown dogs are. It is important to stress, however, that on Fodor's view such entailments are licensed by the *logical structure* of [BROWN(x) & DOG(x)].

[9] Fodor (1998: 44) states that "BROWN DOG is the concept whose extension is the *intersection* of the brown things with the dog things." [my emphasis]. Notice further that Fodor takes properties to be *sets* rather than universals. For example, he writes (2008: 141) "The *extension* of a concept is the *set* of (actual or possible) instances of the property to which the concept is locked. The concept DOG is locked to the property of being a dog and its *extension* is thus the *set* of (actual or possible) dogs." [my emphases].

Contrapositively, if *atomic* concepts have no constituent structure then they have no decompositional analysis and thus license no such entailments. And this is again to deny that putative definitions of basic concepts such as KILL(x) = CAUSE-TO-DIE(x) and inferences such as DOG(x) ⊃ ANIMAL(x) are *analytic*.

3.3. Concepts and prototypes

While Fodor denies that basic/primitive concepts are *constituted* by their prototypes, he does not deny the psychological reality of prototypes. Indeed, Fodor grants that acquiring a prototype may well play a causal role in concept acquisition; not as a matter of nomological necessity mind you, but simply as a matter of brute empirical fact.[10] For on anyone's account, the acquisition of at least some basic concepts in some way depends on experiences with their prototypical instances. And since prototype formation is just a natural consequence of having had such experiences, Fodor grants that concept acquisition may often occur in consequence of having first formed a prototype of the concept to be acquired. However, Fodor's acquiescence on this point is conditioned upon a particular conception of (i) the nature of prototypes, (ii) the role of experience in prototype formation, and in turn (iii) the role of prototypes in concept acquisition.

So before going further, it will help to first clarify what prototypes are commonly supposed to be. As a signpost, my aim in trying to get clear about the role of prototypes in concept acquisition will aid in the construction of a working hypothesis about how concepts$_W$ are acquired. And in the next chapter I argue that the construction of

[10] Fodor (2008: 150) states that "Although it's quite true that acquiring a concept can't be the same thing as learning its stereotype, it needn't follow that learning a stereotype is just a *by-product* of acquiring the concept; it could rather be a *stage* in concept acquisition. Or, to put the suggestion the other way around, concept acquisition might proceed from stereotype formation to concept attainment."

prototypes for other yet-to-be-acquired lexicalizable concepts may initially involve the use of concepts$_W$.

3.3.1. What are prototypes?

As commonly characterized, prototypes are complex mental representations whose representational contents are statistical abstractions over properties instantiated by prototypical members of a given ontological category (i.e., class of things that are presumed to exist in the world). More specifically, prototypes are often described in the literature as lists or sets of semantic "attributes" or "features" each representing a particular property that individual members of that category are statistically *likely* to possess.[11] Or as Margolis & Laurence (1999: 27) put it, prototypes "are complex representations whose structure encodes a *statistical summary* of the properties their members *tend* to have."[12] This tendency is often measured as a function of the relative frequency of occurrence of these properties among instances of the category. For instance, Murphy (2002: 45-6) summarizes the notion thusly:

> The view taken by Rosch and Mervis (1975), Smith and Medin (1981), and Hampton (1979) was that the category representation should keep track of how often features occurred in category members. For example, people would be expected to know that

[11] In this context, conceptual features are also sometimes referred to as 'attributes'. However, recall from Chapter 2 that this use of the term 'feature' is not to be confused with Chomsky's notion of linguistic features.

[12] Given failed attempts to analyze lexical concepts in terms of necessary and sufficient conditions on class membership, most prototype theorists nowadays adopt the weaker stance that prototypes reflect a *statistical summary* of properties instantiated by prototypical instances of the class (*cf*., Murphy [2002: 42]). Subjects then apply these prototypes with varying degrees of success, based on some sort of similarity metric, when judging whether a particular individual is an instance of the class that the prototype subsumes. Margolis & Laurence (*ibid*: 28) point out, however, "that the term "prototype" doesn't have a fixed meaning in the present literature and that it's often used to refer to the exemplar that has the highest typicality ratings for a superordinate concept…"

> ‘‘fur’’ is a frequent property of bears, ‘‘white’’ is a less frequent property, ‘‘has claws’’ is very frequent, ‘‘eats garbage’’ of only moderate frequency, and so on.[13]

To a first approximation, then, we might think of prototypes as lists of semantic features that represent prototypical qualities of their instances, and for each feature an assigned weight corresponding to its relative frequency of occurrence among all instances of the category.

One question that immediately arises about prototypes so-described, however, has to do with the exact nature of the features/attributes that constitute them. Specifically, one wants to know whether prototype features are *concepts* or whether they are perhaps *non-conceptual* in nature. From what I gather, prototype theorists typically take features to be statistically qualified concepts. To call them concepts is to say prototype features represent full-blown ontological categories that thinkers consciously apply to their experiences of the world as a way of carving those experiences at their metaphysical joints. By contrast, to say that features are “statistically qualified” is, in the general case, just to say that not all instances of the category necessarily instantiate each of the prototype’s encoded features, although some features may be perceived as (and may in fact be) essential to the category.

To take a toy example, let us suppose that prototypical dogs are furry, four-legged animals. That is, by assumption anything that is a dog probably instantiates the ostensible properties of being furry and being four-legged and, one assumes, necessarily instantiates the species-essential property of being an animal. For simplicity, let us also just stipulate that highly variable/contingent properties such as being a particular color, weight, age,

[13] Murphy himself eschews a “feature list” view of prototypes in favor of what he calls “schemata” to represent semantic primitives/features. So far as I can tell, however, the “feature list” view remains the predominant view of prototypes among cognitive psychologists.

and so forth are non-constitutive of the dog-prototype. Granted the assumption that features are concepts, a typical thinker's dog-prototype will therefore typically list concepts such as FURRY(x), QUADRUPED(x), and ANIMAL(x). In concrete terms, let us assume that the traditional Roschian prototype for dogs is specifiable along the lines of DOG_P,

$$\text{DOG}_P = \text{x: } \{\{\text{FURRY(x)}, w_1\}, \{\text{QUADRUPED(x)}, w_2\}, \{\text{ANIMAL(x)}, w_3\}\}$$

where w_1, w_2, and w_3 reflect the relative weights of the named features/properties, again in terms of their frequency of occurrence, or perhaps merely their perceived centrality to the category. And we can think of 'DOG_P' merely as an index to the dog prototype stored in a subjects long-term conceptual-semantic memory.

If the constituents of prototypes are concepts then prototype theorists must allow for the possibility that constituents of DOG_P are themselves complex, which is to say decomposable into further features/concepts that may have their own prototypes. For instance, the concept ANIMAL(x) is *prima facie* describable by a prototype whose constituents include the more basic concepts ORGANISM(x) and MULTICELLULAR(x). These latter concepts are presumably further decomposable into still more basic features until the analysis eventually bottoms out in some small set of semantic "primitives," which is to say *sui generis* concepts such as CAUSE, EVENT, OBJECT, SUBSTANCE, and so on. The relevant point, however, is that if basic concepts are *constituted* by their prototypes, as prototype theorists suppose, then DOG_P (or something near enough) is putatively what typical thinkers use to represent dogs *as such*, which is to say *as dogs*.

3.3.2. Why prototypes can't be concepts

Now, if Fodor's is correct in thinking that concepts compose in predictable ways (i.e., productively and systematically), and prototypes are supposed to be concepts, then one would expect prototypes to compose in predictable ways. However, recall that Fodor's chief objection to prototypes *as a theory of concepts* is that they do not compose in predictable ways.[14] To invoke Fodor's favorite foil, while one might agree that prototypical brown dogs are both prototypically brown and perhaps prototypically dog-like, prototypical pet fish, say goldfish or guppies, are neither prototypical pets nor prototypical fish. And this is to say that the prototype for pet fish is not, as prototype theory predicts, a composite prototype such as [PET$_P$ & FISH$_P$].[15] Rather, [PET$_P$ & FISH$_P$], if such a representation is even constructible by human minds, would presumably describe an impossible creature such as a *furry four-legged trout*![16] Or at any rate, it has yet to be shown how prototypes compose to deliver the expected results in all cases (though see Smith & Osheron [1984] for a serious attempt). Moreover, according to Fodor the reason that prototypes don't compose is that although they are complex mental representations, understood as mere lists or sets of features prototypes lack the kind of *logical form/structure* required of linguistic compositionality.

The relevant consequence is again that if concepts compose but prototypes do not then prototypes cannot be concepts. In Fodor's words (1998: 94):

> In a nutshell, the trouble with prototypes is this. Concepts are productive and systematic. Since compositionality is what explains systematicity and productivity, it must be that

[14] It may of course be possible to demonstrate that prototypes can combine using, say, set-theoretic operations such union and intersection. However, Fodor's point seems to be that these modes of combination do not comport with the way human concepts compose.
[15] *Cf.*, Fodor & Lepore (1996) for discussion.
[16] I think even Fodor grants that synthetic compounds such as 'catfish' may have atomic yet idiomatic meanings (express idiomatic atomic concepts).

> concepts are compositional. But it's as certain as anything ever gets in cognitive science that prototypes don't compose. So it's as certain as anything ever gets in cognitive science that concepts can't *be* prototypes and that the glue that holds concepts together can't be statistical.

Be that as it may, Fodor again acknowledges the psychological reality of prototypes. In particular, he is responsive to the evidence of so-called "typicality effects" adduced to prototypes regarding our judgments about category membership, which is what prompted Eleanor Rosch and her cohorts in the 1970's to first hypothesize that concepts are defined by their prototypes (or stereotypes—Fodor uses these terms interchangeably).[17] What's more, as stressed earlier Fodor also concedes that in the general case prototype formation quite likely constitutes a stage in concept acquisition (more on this below).[18]

Unfortunately, nowhere does Fodor say exactly what he takes prototypes to be, except in the highly general sense of being "the estimation of central tendencies" or "abstractions from experience." To insist, as Fodor does, that prototypes are not concepts can however be interpreted in at least two ways. In the first instance, we can take him as saying that prototypes are *non-conceptual* mental representations. And as discussed below there is reason to think that this is in fact what Fodor in fact has in mind. In the second instance, Fodor might merely mean that prototypes cannot be *basic/primitive* (i.e., unanalyzable) concepts, which leaves open the possibility that at least some prototypes are *complex* concepts, yet under the following provisos: (i) that their constituents are themselves concepts that subjects already possess, and (ii) that their contents are

[17]E.g., judgments such as that robins are more bird-like (more typical examples of the category BIRD) than, say, hummingbirds, ostriches, and penguins.

[18] Specifically, he writes (2008: 150, fn.21-2): There's convincing experimental evidence that something like the estimation of central tendencies does go on in stereotype formation… At a minimum, there is quite a lot of empirical evidence that stereotype formation often happens *very early* in concept acquisition; young children's judgments about what concepts apply to what things are typically most reliable for paradigm instances. They're good at whether dogs are animals long before they've got the extension of ANIMAL under control.

compositionally-determined like any other complex concept. While the latter seems not to be Fodor's considered view, it strikes me as compatible with it, and carries certain benefits. Indeed, I think there is room for both conceptions of prototypes within a Fodorian theory of concepts, and I will return below to motivate this hypothesis. But it will help to facilitate that discussion by first briefly reviewing Fodor's expressed account of concept acquisition, drawing chiefly from Fodor (1998) which so far as I can tell reflects his most deeply considered thought on the matter.

3.4. On concept acquisition

To repeat, acquiring a Fodorian concept is getting nomologically locked to the property that the concept expresses, whereas concept possession entails being so-locked to that property. Now, Fodor again concedes that "nobody actually knows what 'exogenous' variables may affect a creature's capacity for conceptualization," including (as he quips) diet, being hit on the head by a brick, contracting senile dementia, and moving to California (the latter, I take it, is intended as a jab at Connectionists). In any case, these are all nomological possibilities. However, Fodor again grants that it may be a brute empirical fact that concept acquisition occurs in consequence of having had the right kinds of experiences with their prototypical instances. In other cases, as discussed below, getting locked to a property involves learning and adhering to a (true) *scientific theory* of the property in question.

More specifically, Fodor maintains that there are basically two kinds of properties that minds like ours can become nomically locked to, corresponding to two ways of being locked to them. There are what he calls *mind-dependent* properties and those that are *mind-independent*. As the name suggests, mind-*dependent* properties, or what I will call

MDPs for short, depend for their existence on the psychological laws that govern how things that have them appear to us. Or stronger still, as Fodor puts it, MDPs are *constituted* by psychological laws that govern the way their instances "strike our kind of minds as being." Paradigm examples of MDPs are sensory properties such as *being red*, which Fodor claims are "constituted by the sensory states that things that have it evoke in us." In turn, it is visual experiences of things "as-of" being red that causally mediate the acquisition of the sensory concept RED(x). Notice again, however, that to say that concept acquisition is causally mediated by experience is not to say that concepts are learned by statistical abstraction over these experiences (though the *prototypes* that mediate their acquisition may be learned by statistical abstraction). Rather, in these cases Fodor likens concept acquisition to contracting the flu—it is something that just happens to us *in consequence* of having the right kinds of experiences.[19]

Sensory properties are members of a more general category of MDPs called *appearance properties* which we get locked to by way of *appearance concepts* including our old friends DOORKNOB(x) and CARBURETOR(x). Like sensory properties, appearance properties are constituted by a psychological law that governs how "we lock to them in consequence of certain sorts of experience," and in particular experience with things that exemplify their prototypes. And this explains why the acquisition of appearance concepts is typically mediated by the formation of a prototype. Fodor sums up the distinction as follows:[20]

[19] This raises the interesting question of whether blind people can acquire sensory concepts. For a proposal in the affirmative, see Zalta (2001).

[20] Recall that Fodor is a realist about colors, doorknobs, and carburetors; to be mind-dependent does not imply being unreal. Yet by the same token if there were no minds there would be no colors, carburetors, or doorknobs, or for that matter any other appearance property.

> The model, to repeat, is *being red*: all that's required for us to get locked to *redness* is that red things should reliably seem to us as they do, in fact, reliably seem to the visually unimpaired. Correspondingly, all that needs to be innate for RED to be acquired is whatever the mechanisms are that determine that red things strike us as they do; which is to say that all that needs to be innate is the sensorium. Ditto, *mutatis mutandis*, for DOORKNOB if *being a doorknob* is like *being red*: what has to be innately given to get us locked to *doorknobhood* is whatever mechanisms are required for doorknobs to come to strike us as such.

The challenge is of course to spell out more precisely the nature of the cognitive and/or neurological mechanisms that are responsible for the way that objects and their properties "strike our kind of minds as being," which I return to momentarily.

In contrast to MDPs, mind-*independent* properties, or what I will call MIPs for short, exist independently of minds—they are *not* properties that things have "in virtue of their relations to minds, ours or any others." For acquiring a concept of an MIP involves getting locked to a property via a nomic mind-world relation. Paradigm examples of MIPs are natural kinds—properties that constitute the "hidden essences" of things that are causally responsible for their superficial qualities/appearances.[21] Rather than being constituted by psychological laws, MIPs are governed by laws invoked by the physical sciences. The chemical essence of water, for example, is H2O, which is presumably what determines its superficial qualities. However, according to Fodor there are actually two ways of getting locked to the property of *being water*, which is to say the property of *being H2O*. The first (and most common) is what I will call the pre-theoretical way, and the other is by way of constructing a formal scientific theory of water.

[21] Fodor suggests that logico-mathematical properties are also mind-independent, but I'll set aside this question.

In advance of formal training in chemistry and physics, most of us get locked to water via acquisition of a pre-theoretical concept that I will call WATER(x)^{PT}. The acquisition of WATER(x)^{PT} proceeds in roughly the same way as the acquisition of appearance concepts such as DOORKNOB(x), which is to say in consequence of experiences with its prototypical instances. The relevant difference is that being locked to a natural kind such as water, *viz*. the property of *being H2O*, by way of a pre-theoretical concept "is reliable only in worlds where water *has* the familiar phenomenological properties; which is to say only in nomologically possible worlds near ours." To again quote Fodor at length:

> That is, I suppose, the usual, pretheoretic way of having a natural kind concept. The kind-constituting property is a hidden essence and you get locked to it via phenomenological properties the having of which is (roughly) nomologically necessary and sufficient for something to instantiate the kind… WATER, like DOORKNOB, is typically learned from its instances; but that's not, of course, because *being water* is mind-dependent. Rather, it's because you typically lock to *being water* via its superficial signs; and, in point of nomological necessity, water samples are the only things around in which those superficial signs inhere.

In short, Fodor adds that "A 'natural kind concept' can be [*merely*] the concept of a natural kind; or it can be the concept of a natural kind *as such*…"

By Fodor's lights, the only way to get locked to water *as such*, which is to say the property that water has in every metaphysically possible world, is by learning a (true) scientific theory of water *qua* H2O. On this count he qualifies:

> We're locked to water via a theory that specifies its essence, so we're locked to water in every metaphysically possible world. That, I'm suggesting, is what an informational semanticist should say that it *is* to have a concept of a natural kind *as* a natural kind: it's for the mechanism that effects the locking not to depend on the superficial signs of the

> kind, and hence to hold (*ceteris paribus* of course) even in possible worlds where members of the kind lacks those signs.

As it turns out, there are also two ways of getting theoretically locked to water *as such*. The first is by acquiring and then combining more basic concepts such as HYDROGEN(x), TWO(x), and OXYGEN(x) to form the complex concept H2O(x), or derivatively a complex concept such as [HYDROGEN(x) & DIOXIDE(x)], whose compositionally-determined content is the property of *being H2O* (or *being hydrogen+dioxide*). However, Fodor suggests that one can also get theoretically locked to water *atomistically* by way of what he calls "the *real* concept WATER," or what I will call the scientific concept WATER(x)ST. The difference here is that WATER(x)ST is acquired in virtue of learning, or inventing, and otherwise adhering to a (true) scientific theory of water's hidden essence (whose construction presupposes possession of the concepts HYDROGEN(x), TWO(x), and OXYGEN(x)). As Fodor (1998: 162) puts it:

> A theoretical concept isn't a concept that's *defined* by a theory; it's just a concept that is, de facto, locked to a property via a theory.

That's the long and short of it. To recap, thinkers typically get locked to MDPs by way of experiences with their prototypical instances, whereas getting locked to a MIP occurs *either* by experience *or* by learning/inventing[22] a (true) scientific theory of its "hidden essence" (i.e., a theory that establishes the empirical truth of the statement 'water = H2O'). That the theory be *true* is important, as evidently concepts grounded in false theories, e.g., PHLOGISTON(x), VULCAN, etc., never lock to a MIP *as such*. Nor, obviously, can one acquire these concepts through experiences with their prototypical instances (for there are none). Nonetheless, Fodor seems to be a realist about things like phlogiston and

[22] Fodor suggests that acquisition of scientific concepts can also be mediated by instruments such as telescopes and also by deference to experts.

Vulcan. In such cases, I take it that misguided scientists acquire something like a theoretical analog of an appearance concept—a concept that locks to a mind-*dependent* property by way of a *false* theory of how the world is supposed to be. Alternatively, it may be that some of our mind-dependent concepts are locked to properties that simply have no instances in the actual world. In a similar vein, purely fictional concepts such as UNICORN(x) get locked to (uninstantiated) properties by way of a story, or myth, or what have you.[23] It's just that in the latter case the concepts are often acquired without existential pretense. In this way, some concepts are acquirable by description rather than direct acquaintance with their instances.

3.4.1. On the role of prototypes in concept acquisition

So characterized, let me now return to the question of how prototypes and/or scientific theories stand to mediate the acquisition of Fodorian atoms—i.e., basic, unanalyzable concepts. As mentioned earlier, to say that prototypes are not concepts can be interpreted in at least two ways. First, we might take Fodor as implying that prototypes are non-conceptual in nature. Second, Fodor might merely mean that prototypes cannot be basic/atomic concepts, which leaves open the possibility that prototypes are complex concepts (e.g., beliefs). For my part, I think there is room, with suitable qualification, for both conceptions of prototypes within a Fodorian theory of concepts.

In what follows immediately, I will offer my best reconstruction of how prototypes might be recast as non-conceptual (or pre-conceptual) mental representations, or what I will call *F-prototypes* in honor of Fodor. I will agree with Fodor that *F-prototypes*

[23] As Fodor would put it, while the property of being a unicorn has not instances in the actual, it might in some nearby possible world. That is, counterfactual unicorns do exist and thus so does the property of being a unicorn.

facilitate the acquisition of some/many atomic concepts, and in particular those that he refers to as "sensory concepts." I will then attempt to make a case for higher-order conceptualized prototypes, or *C-prototypes* as I shall call tem, that by hypothesis facilitate the acquisition of certain human concepts such as JUSTICE, LOVE, and PET-FISH. For notice that from the claim that prototypes do not compose in the way that concepts do it does not follow that prototypes do not compose. I will further argue C-prototypes are needed in any case to account for Roschian typicality effects regarding our explicit judgments about category membership.

Let me pause a moment to remind readers that I am ultimately working toward a psychologically plausible account of the nature of concepts$_W$, how *they* are acquired, and whose contents, to foreshadow, I take to be *linguistic* (i.e., *scientific*) *kinds*. As such, I will later argue alongside Fodor that there are two ways of getting locked to Words (i.e., the property of being a certain *lexical item* in a speaker's mental lexicon). First, there is the ordinary way of acquiring a pre-theoretical concept$_W$ of the Word in question. The second is by learning/formulating a theory (such as Chomsky's) about the underlying psychological/scientific nature of Words.

3.4.2. On the nature of F-prototypes

To begin, it bears noting that Fodor's informational atomism is modeled after the Dretskian idea that a naturalized theory of reference can be given in terms of the information carried by token mental representations about their distal causes. Specifically, Fodor (2008: 179) suggests that reference can be characterized by the schema: "X represents [or refers to] Y insofar as X carries information about Y." With respect to early vision, for example, the idea is that patterns of retinal stimulation carry

information about their distal causes—e.g., objects and their properties in a subject's occurrent field of vision. This information then serves as the raw, unconceptualized (or pre-conceptual) content of our visual experiences—i.e., that which is perceptually "given" in experience, and which as Fodor reports is standardly taken to be *iconic* (as opposed to *discursive*) in nature.

What is given in perception is in turn what grounds the fixation of our fully conceptualized perceptual judgments/beliefs.[24] As Fodor (2008: 180-2) puts it:

> The idea is that the role of concepts in the perceptual analysis of experience is to recover from experience information that it contains… The given is unconceptualized representation that is awaiting conceptualization.

He adds (*ibid*: 193), however, that:

> There is every sort of evidence that a great deal of the reasoning involved in the causal fixation of quotidian perceptual beliefs is unconscious and hence unavailable for report by the reasoned…

In other words, the early visual system registers information without subjects perceiving this information as representing anything in particular. Furthermore, "what information an interpreter can recover from [a given perceptual experience] depends on what concepts the interpreter has available." And so in this way "the notion of carrying information about X seems to offer a way of representing X without representing it as anything…" (*ibid*: 182).

More specifically still, Fodor flirts with the idea that perceptual reference is grounded in something like the FINST mechanism proposed by Pylyshyn (2003, 2009).

[24] Fodor (2008: 185) summarizes the idea thusly: "The basic idea is that perceptual information undergoes several sorts of process (typically in more or less serial order) in the course of its progress from representation on the surface of a transducer (e.g. on the retina) to its representation in long-term memory. Some of the earliest of these processes operate on representations that are stored in an 'echoic buffer' (EB), and these representations are widely believed to be iconic."

By hypothesis, FINSTs (for "fingers of instantiation") operate in early vision as primitive referential indices that lock onto and track perceptually salient distal objects in a subject's visual field. This "locking onto" is described as an automatic (i.e., purely psychophysical/causal-mechanical) reflex to the proximal retinal image (or "FING") that is itself caused by the distal object perceived. Or as Pylyshyn puts it, FINSTs are "grabbed" by FINGs. Similar to Fodor's view of conceptual reference, this relation is referentially primitive in that the locking mechanism is grounded in a nomological link between FINSTs and their proximal causes. And so in this way Pylyshyn's FINST mechanism comports with Fodor's informational atomism about concepts. However, unlike Fodorian concepts, FINSTs are by hypothesis non-conceptual (or again *pre*-conceptual) mental representations—they represent distal objects and their properties without representing them *as such*, which is to say without subsuming their referents under corresponding concepts.

A second difference between FINSTs and Fodorian concepts is that the referential link between FINSTs and their referents is *transient*. Specifically, a limited number of FINSTs (roughly four) are dynamically allocated to perceptibly distinct objects in the perceiver's visual field. FINSTs will then continue to track those objects through space and time, tolerating a certain degree of change in location, trajectory, color, shape, and even momentary occlusion, so long as the causal link between them remains largely uninterrupted. Otherwise, an object must be re-acquired, perhaps under a different FINST. In this way, the brute causal mechanism underlying our visual experiences is loosely analogous to the natural mechanism that causes sunflowers to "blindly" track the sun across the daytime sky. For when the sun disappears over the horizon, or heavy cloud

cover rolls in, the causal-referential link is severed and the tracking ceases. The chief difference here is that *ex hypothesi* FINSTs are *representational* mechanisms whereas sunflowers presumably are not.

More to the point, the claim, as I understand it, is that once locked to an object FINSTs provide a channel that facilitates extraction of information about that object's perceptible properties. Property information is then recorded in what's called an "object file" held in short-term memory, or roughly speaking what memory scientists call an "echoic buffer." From a computational perspective, property information accumulates in what we might think of as a set of general purpose registers corresponding to what Pylyshyn refers to as "visual predicates," one for each detectible property of the object being tracked.

By way of example, let f_1 be a FINST locked to a particular object, say Fido, in the subject's field of view. For each discernible property of that object, early vision allocates a corresponding visual predicate, the set of which I will represent as x: $\{P_1(x), P_2(x), P_3(x), \ldots, P_n(x)\}$. FINSTs themselves then serve as arguments to these (non-conceptual) predicates. For instance, let us suppose that Fido is mostly brown in color. Let us further suppose that P_1 is dedicated to representing object-color. The "saturated" visual predicate $P_1(f_1)$ then demonstratively represents the object referred to by f_1 as instantiating the color property picked out by P_1, *viz.*, the property of *being brown*. The remaining predicates $\{P_2(\mathrm{x}), P_3(\mathrm{x}), \ldots, P_n(\mathrm{x})\}$ track other visually detectible properties related to, say, the object's texture, shape, size, and so forth. Collectively, the information encoded by f_1:

$\{P_1(f_1), P_2(f_1), P_3(f_1), \ldots, P_n(f_1)\}$ represents the raw, unconceptualized informational content of a perceiver's visual experience of the object tracked by the FINST labeled f_1.[25]

It bears stressing here that f_1 does not represent its object (i.e., Fido) *as* a dog, nor does P_1 represent that object *as* being brown. To put the point differently, f_1 is not content-identical with the Fodorian concept DOG(x) (nor FIDO), and P_1 is not content-identical with the Fodorian atom BROWN(x). Likewise, the saturated predicate $P_1(f_1)$ is not content-identical to the complex concept [BROWN(x) & DOG(x)]. Rather, in order to represent an object as a brown dog requires first *attending to* the information encoded by $P_1(f_1)$—the raw content of one's visual experience—and then *applying* appropriate concepts to the content of that experience. However, the lack of corresponding concepts is no barrier to perceptually representing brown dogs, as it were, *de re*. Or as Fodor (2008: 186) puts the point "it's possible to register the Dretskian information that *a* is F even if you don't have the concept F." It is important to keep in mind, however, that without the appropriate concepts one cannot represent brown dogs *as such*, which is again to say *as brown dogs*.

Most relevant to my purposes here is that (if I interpret Fodor correctly) *prototypes* of visually perceptible objects and their properties are formed in consequence of experiences registered in early vision (a similar story can be told for other sensory modalities). Specifically, I take it that the Fodorian prototype, or what I will again abbreviate as an *F-prototype*, corresponding to the concept DOG(x) can be specified as something like DOG_P',

$$\text{DOG}_P' = f_n\text{: } \{\{P_1(f_n), w_1\}, \{P_2(f_n), w_2\}, \{P_3(f_n), w_3\}, \ldots\}$$

[25] In cases of multiple object tracking, something similar can be said of the FINSTs designated f_2, f_3, and so on for each object that can be simultaneously tracked in early vision.

where f_n represents an abstraction over dog-experiences as captured by FINSTs over the course of multiple encounters with "good examples" of dogs, and the pre-conceptual predicates $P_1(f_n)$, $P_2(f_n)$, $P_3(f_n)$, …, correspond to tacitly inferred abstractions over the prototypical properties of dogs extracted from those experiences (the properties of being brown, furry, quadrupedal, or what have you). As before, w_1, w_2, and w_3 represent relative weights assigned to these properties. So characterized, it will be noticed that DOG_P' is structurally identical to the more traditional prototype DOG_P from above, or what I will call an *R-prototype* ('R' for Roschian); the difference being that the informational content of DOG_P' is wholly *non-conceptual* in nature.

Let me be clear that Fodor does not explicitly endorse any of what I have just said about the nature of F-prototypes. Rather, the picture sketched is my best *reconstruction* of how prototype formation might plausibly occur, and how F-prototypes in particular are represented, as based on Fodor's reported account of non-conceptual content. Yet more needs to be said. For according to Fodor what is "given" in perception is registered in the echoic buffer (EB), which is a highly ephemeral, short-term memory register. In consequence, introspective access to the content of visual experience is highly time-sensitive. As Fodor describes it:

> …there is a brief interval during which an unconceptualized (presumably iconic) representation… is held in the EB… it lasts only for perhaps a second or two… After that, the trace decays and you've lost your chance.

Now, if the content of a visual experience is short-lived then one wonders how perceivers use this information to form prototypes. For the computation of a prototype presumably requires access to *accumulated* (i.e., *stored*) memories of *multiple experiences*, both past and present, over which the relevant abstractions/generalizations are inferred.

While Fodor is again silent on such questions, one is given to assume that mental images of visual experiences are recorded in some sort of long-term iconic memory. Thus, it would be these sorts of iconic memories over which "central tendencies" are computed. It is of course quite possible that I have completely misunderstood Fodor's view on the matter. Yet it's also difficult to imagine what else it might be. In any case, I'm happy to assume that subjects are capable of forming primitive, non-conceptual prototypes—or again *F-prototypes*. More specifically, I will assume that F-prototype formation can occur as the result of a subject's tacit registration of patterns of similarity in the detectible properties instantiated by members of a given category.

This datum nevertheless leaves open the questions of (i) how F-prototypes facilitate the acquisition of basic concepts, and (ii) how appeal to F-prototypes might explain Roschian typicality effects as reflected in our explicit judgments about category membership. In the first instance, Fodor's decision to construe prototypes as non-conceptual appears to be motivated by his commitment that concept acquisition is necessarily non-inferential. However, he allows that even the encapsulated processes of early vision facilitate "subpersonal inferences" over the informational content of visual experience, which reportedly explains perceptual constancies such as "the elliptical plate that looks round, the 'correction' of perceived color for changes in the ambient light, the failure of retinal size to predict apparent size, and so forth through a very long list of familiar examples" (Fodor [2008: 192]). On this point Fodor adds (*ibid*):

> …all varieties of CTM treat the perceptual constancies as paradigm examples of the products of subpersonal inferences; that is, they imply that, invariably, mental representations that exhibit the effects of constancy are inferred.[26]

As the outcome of this particular question is largely orthogonal to my project, I propose to set it aside. Let us instead consider the possibility of another kind of prototype grounded in basic concepts, or what I introduced above as *C-prototypes*, which are to be distinguished from both F-prototypes and R-prototypes.

3.4.3. On the nature of C-prototypes

While one might grant that F-prototypes causally mediate the acquisition of sensory concepts such as BROWN(x) and FURRY(x), it remains unclear (at least to me) how F-prototypes account for the acquisition of concepts that are, as it were, less intimately tied to the sensorium, including I assume most quotidian concepts like DOORKNOB(x), CARBURETOR(x), and PET-FISH(x). It is also unclear how F-prototypes are supposed to explain Roschian typicality effects related to our explicit judgments about category membership. There is however a way of accommodating these *explananda* without assuming that *all* prototypes are non-conceptual in nature while also keeping to the spirit of Fodor's account of basic concepts. Specifically, the proposal I wish to pursue is that while sensory concepts such as RED(x) may be acquired in virtue of acquiring an F-prototype, many other basic concepts are acquired in virtue of learning/constructing *C-prototypes*, which is to say complex concepts whose compositionally-determined contents are criterial of their instances.

[26] However, *prima facie* this claim stands in direct contradiction to an earlier remark (*ibid*: 185) stating that "Inferring is in the same basket with as saying and thinking; they all presume conceptualization."

3.4.4. Supplementing F-prototypes with C-prototypes

We have seen, for example, that on Fodor's view anyone already in possession of the basic concepts FURRY(x), QUADRUPED(x), and ANIMAL(x) can *use* these concepts to construct the compositionally-determined complex concept [FURRY(x) & QUADRUPED(x) & ANIMAL(x)]. Now, under the caricature assumption that [FURRY(x) & QUADRUPED(x) & (ANIMAL(x)] is criterial of prototypical dogs, one might hypothesize that rather than DOG$_P$ or DOG$_P$' from above, the prototype for dogs is instead more like DOG$_P$" below:

DOG$_P$" = x: [(FURRY(x), w_1) & (QUADRUPED(x), w_2) & (ANIMAL(x), w_3)]

where w_1, w_2, and w_3, as before, reflect the relative weights of the named features. But in the present example, I propose that we think of these weights as independent parameters that "tag along for the ride" with respect to compositional processes.[27] Loosely speaking, then, we can think of DOG$_P$" as reflecting a subject's belief (or theory, or hypothesis) that there exists a class of entities in world—the set of dogs—whose members probably instantiate the property of being a furry four-legged animal. Now, I assume with Fodor that weights don't compose in the relevant way. And so *strictly speaking* DOG$_P$" is not a *concept* in Fodor's sense of the term. This constraint notwithstanding, I see no reason to deny that DOG$_P$" is representable by human minds and can thereby serve as a conceptually grounded prototype (or again C-prototype) for the concept DOG(x). Or anyhow, I will assume for purposes here that it can be.

On this construal, C-prototypes can be learned like any other complex concept simply by conjoining more basic concepts, and indeed is what makes learning complex concepts a *rational* achievement. Importantly, however, whether or not DOG$_P$" is a

[27] Although in combination, the "inherited" default weights assigned to individual prototypical properties of basic concepts may require adjustment depending on which other concepts enter into the combination.

concept, strictly speaking, by Fodor's criteria one still cannot use DOG$_P$" to think about dogs as such. Rather, for this one must first be in possession of the conceptual *atom* DOG(x). In other words, I am here simply indicating agreement with Fodor that DOG$_P$" and DOG(x) are not content-identical. For the concept [FURRY(x) & QUADRUPED(x) & ANIMAL(x)] applies only to prototypical dogs, whereas *ex hypothesi* DOG(x) ranges over all dogs, including atypical ones. And this is just to recapitulate Fodor's more general observation that the contents of basic concepts such as DOG(x) outrun our experiences of their prototypical instances. Nor again can [FURRY(x) & QUADRUPED(x) & ANIMAL(x)] be *constitutive* of DOG(x). For one can have [FURRY(x) & QUADRUPED(x) & ANIMAL(x)] without DOG(x), and *vice versa*.[28]

Thus qualified, it seems to me entirely compatible with Fodor's view that learning DOG$_P$" can nonetheless causally mediate the acquisition of DOG(x) while allowing that its basic constituents are grounded in F-prototypes. Otherwise, it remains unclear what information obtained from the sensorium guides a thinker's locking onto the property of being a dog, or for that matter being a carburetor, being a doorknob, being a pet, being a fish, and so on. By contrast, according to the proposal on offer C-prototype formation is akin to forming a *theory*, albeit informal/unscientific in ordinary cases, about which properties individual members of a given category are likely to share. However, I find it more useful to think of C-prototypes—or anyhow purely discursive ones—in terms of *beliefs*, or sets of *belief-predicates*, about which properties members of given class are likely to have in common. For what is a theory if not a set of belief-statements?

[28] In the other direction, Fodor's auxiliary thesis of "reverse compositionality" explains why one cannot believe that dogs, *as such*, are furry four-legged animals in advance of having the concept DOG(x).

Furthermore, I assume that once a subject has acquired DOG(x), he/she will be in a position to re-formulate DOG_P" as something like the following universally quantified belief:

$$\text{DOG}_P''' = \forall x\,[\text{DOG}(x) \supset \text{P-INSTANTIATES}(x, (\text{FURRY}(x), w_1)\ \&\ (\text{QUADRUPED}(x), w_2)\ \&\ (\text{ANIMAL}(x), w_3)])]$$

where the concept P-INSTANTIATES(x, Y) reflects an agent's belief that prototypical members of the category DOG(x) probably instantiate the properties *furry, quadruped,* and *animal*.

Returning to a previous example, it is similarly compatible with Fodor's view that anyone who has the concepts PET(x) and FISH(x) can use these concepts to form the complex concept [PET(x) & FISH(x)]. And anyone who has [PET(x) & FISH(x)] can use this concept to generate beliefs about *whatever* one's conception of prototypical pet fish happens to be. *To wit*, I think of pet fish as being exotic, as dwelling in fish tanks, as being short-lived, etc., which is to suggest that my conception (*C-prototype*) of pet fish can be specified as follows:[29]

$$\text{PET-FISH}_P = \forall x\,([\text{PET}(x)\ \&\ \text{FISH}(x)] \supset \text{P-INSTANTIATES}(x, [\text{EXOTIC}(x)\ \&\ \text{INHABITS-FISH-TANKS}(x)\ \&\ \text{SHORT-LIVED}(x)])$$

In other words, I take it that PET-FISH_P is the representation I deploy when judging whether or not something qualifies as a pet fish. Relatedly, I have in addition to the belief above the one below,

$$\text{DEAD-PET-FISH}_P = \forall x\,([\text{PET}(x)\ \&\ \text{FISH}(x)]\ \&\ \text{DEAD}(x) \supset \text{P-INSTANTIATES}(x, [\text{FLUSHED-DOWN-TOILETS}(x)])$$

[29] I am omitting weighting parameters in the interest of readability, but I assume that they are present in these representations as well.

where [[PET(x) & FISH(x)] & DEAD(x)] expresses the property of *being a dead+[pet+fish]*. Here again, notice that PET(x), FISH(x), and [PET(x) & FISH(x)] are each proper constituents of my beliefs about (or conception of) prototypical pet fish.[30] Indeed, it is a point in Fodor's favor that concepts, whether simple or complex, are in this way fully productive. But notice that all of this can be so without supposing that PET-FISH$_{\boldsymbol{P}}$ (i.e., the *prototype*) expresses the property of *being a pet+fish*, or that DEAD-PET-FISH$_{\boldsymbol{P}}$ expresses the property of *being a dead+[pet+fish]*.

Rather, on Fodor's view the compositionally-determined content of [PET(x) & FISH(x)] is nothing more exotic than the property of *being a pet+fish*. Now, while I have a pretty clear idea of what prototypical pet fish are (as illustrated by PET-FISH$_P$), I haven't much of a clue what it is to be a *pet+fish*. And I take it that neither does Fodor, other than to say that "the set of pet fish is the overlap of the set of pets with the set of fish" (1998: 103). Yet as methodological point, Fodor rightly notes that the metaphysics of conceptual content is in no way beholden to our epistemic relation to properties like *being a pet+fish*. If it turns out that goldfish instantiate *pet+fish* then goldfish are instances of [PET(x) & FISH(x)]. Same goes for guppies, piranhas, catfish, and so on down the trotline. The upshot is that satisfying the possession conditions [PET(x) & FISH(x)] in no way guarantees that its possessors will *recognize* its instances *as such*. In this way, the content of one's C-prototype of pet fish can diverge from the content of the complex concept [PET(x) & FISH(x)]. But more to the point, PET-FISH$_P$ demonstrates how antecedently held concepts can be used to form prototypes constructed from those concepts. It also stands to account for our explicit judgments about category membership.

[30] I imagine that many such beliefs are (or can be) inferred real time, rather than stored in long-term memory, whenever the topic of pet fish arises.

As a point of historical interest, prototype theorists also struggle to explain how two people might ever come to share the same concept. For while a goldfish may be your idea of a prototypical pet fish, mine might be a guppy, or for that matter any species of aquatic creature that I conceive of as being prototypical pet fish. The present account, by contrast, readily explains how two subjects who share the one-and-only concept [PET(x) & FISH(x)] might nonetheless disagree wildly about what counts as a prototypical pet fish. For again it is the contents of their *beliefs* about pet fish and other things that diverge among subjects, and occasionally from reality, not the contents of their constituent concepts. Or as Fodor (2008: 144) argues, an atomistic theory of concepts predicts and thereby explains why, for example, "Ancient Greeks, who thought stars were holes in the fabric of the heavens, could nevertheless think about stars." The basic moral is that having a concept does not depend on the epistemic capacity to identify/discriminate its instances.[31]

I again dwell on such matters because the results can be used to explain how our ordinary *conceptions* of Words and their lexically-encoded meanings can diverge from their actual type-individuating properties. For example, if you ask the average Joe Sixpack on the street what the meaning of a Word is, he will likely fumble through some story to the *effect* that the meaning of a Word is whatever it denotes or refers to.[32] However, if Chomsky is right Words themselves neither denote nor refer in the traditional Russellian sense. Rather, Words are by hypothesis instantiations of the complex mental symbol {PHON, SYN, SEM}. And SEMs, which is again the term used

[31] Fodor (2008: 137, fn.8) adds "That I do not know (and do not know how to find out) whether paramecia are green does not impugn my claim to have GREEN."
[32] More likely I suspect that what you'll get in the case of, say, common nouns is a *description* of things that those Words designate, but the details here are irrelevant to the point being made.

to characterized a Word's lexically-encoded meaning, are purely formal, intrinsic properties of Words that guide and constrain without determining what they can and cannot be used to talk about in ordinary contexts. Rather, if what I have said thus far is correct, it is the *effects* that lexical meanings have on our *use* of Words that informs our ordinary conceptions of them.

More specifically, and as I briefly argued in Chapter 1, I assume that speakers use their *concepts*$_W$ of Words, *inter alia*, to form beliefs about their meanings such that the concept$_W$ in question *ipso facto* becomes a proper constituent of the belief so-formed. The set of such beliefs in turn constitutes a speaker's *conception* of those Words, or given the present discussion what we might think of as their prototypes. Furthermore, as suggested below concepts$_W$ are metaphysically locked, *à la* Fodor, to the same empirical entities that theoretical linguists talk about, which is to say *Words*, which is to say items in a Chomskyan lexicon. That is, when ordinary speakers talk about the meanings of Words I assume that they are talking about (in a *de re* sense) the same things that theoretical linguists talk about when *they* talk about Words. It's just that our ordinary, pre-theoretical conceptions of Words and their lexical meanings fail to map *directly* onto their actual language-internal, type-individuating properties and relations. In other words, while ordinary speakers and trained linguists think of and talk about Words in different ways, one presumes that they are thinking of and talking *about the same things*.

The point, in short, is that having perfectly good concepts$_W$ *of* Words in no way guarantees that their possessors will conceive of Words as being so-constituted, any more than, say, having the concept WATER(x) ensures knowing that water consists in H2O molecules. Rather, to know such facts, as Fodor puts it, requires doing the science. This is

not however to suggest that our ordinary conceptions of Words are systematically mistaken. Rather, the claim again is merely that our pre-theoretical conception of Words and what they mean is to be understood *de dicto*—"ways of thinking about" Words and what they mean as evidenced by certain type-individuating properties of their tokens. And so the claim is that ordinary speakers think about Words and their properties in ways that differ from the ways in which theoretical linguists think about Words. Specifically, I have been arguing that competent speakers know (in a robust sense of "knowing") the various *ways* in which the meanings guide and constrain without determining what their host Words can and cannot be used to talk about in ordinary discourse. And so in this sense, ordinary speakers know perfectly well *what* they are talking about when talking about the meanings of the Words they utter.

3.5. On the nature of concepts$_W$ and their acquisition

I have characterized the contents of concepts$_W$ as expressing a linguistic property; namely, the property of being the Word that they are concepts$_W$ of. For example, I take it that the content of DOG(w)—which is to say one's concept$_W$ of the Word 'dog'—is the property of being the English Word 'dog'. So characterized, however, it is *prima facie* unclear how to classify concepts$_W$ under the Fodorian framework described above. For under present assumptions there is an obvious sense in which concepts$_W$ range over mind-*dependent* properties (i.e., properties of lexical items, understood as *mental representations*). Yet these properties are also presumed to be linguistic (i.e., scientific) kinds, which Fodor characterizes as mind-*independent* properties.

I take the apparent conflict here to be largely terminological. For while Fodor does not specifically entertain the possibility, there is nothing incoherent about treating Words

as natural or scientific kinds, as in fact Chomsky and other like-minded theorists treat them. I doubt that Fodor would disagree. As such, I will assume in what follows that concepts$_W$ are Fodorian atoms whose contents are corresponding items, which is to say representations, in a speaker's mental lexicon. I further assume, as has just been suggested, that ordinary speakers get locked to Words *pre-theoretically* by way of prototype formation, whereas suitably trained linguists get locked to them by way of a *scientific theory* of their intrinsic properties and relations.

In the first instance, I assume that as with WATER(x)^{PT} from above our concepts$_W$ get pre-theoretically locked to Words by way of our perceptual experiences of their prototypical instances—i.e., *tokens* of Words as they occur in verbal discourse. Of course, to have concepts$_W$ of particular Words presupposes having a concept of what Words are, *in general*—call this WORD(w)^{PT}. Similar to our pre-theoretical conceptions of lexical meanings, I take it that our pre-theoretical conceptions of what Words are is likewise blurry. However, if Fodor is right this is no obstacle to acquiring (and thus possessing) the concept WORD(w)^{PT}, and likewise concepts$_W$ of particular Words that are among its instances. For similar to concepts such as DOORKNOB(x) and CARBURETOR(x), what matters is that we perceive Words as being a certain way, and in particular that they strike minds like ours as being Words of natural language.

By contrast, I assume with Fodor that theoretical linguists (*qua* scientists) get locked to Words by way of a formal theory about their hidden essences. For instance, my *scientific concept* WORD(w)ST is currently attached to a theory stating that Words are instances of the mental structure {PHON, SYN, SEM}. Of course, my theory—which is essentially Chomsky's—may turn out to be false, and indeed quite likely is in certain

respects. But under the broader hypothesis that Words are in fact psychologically real linguistic entities, the former deficit is no barrier to getting/being locked to them, or at least no less so than, say, suitably trained physicists are locked to the property of *being a quark*, or that geologists are locked to the property of *being a tectonic+plate*, and cosmologists are locked to the property of *being a black+hole*. Rather, I take it that on Fodor's view having an incomplete or even partially false theory of the property in question just counts as a less robust or perhaps less stable way of getting/being theoretically locked to that property.

In turn, my scientific theory of *particular* Words includes beliefs about specific properties of their lexical entries, including their linguistically-encoded meanings. Now, I have suggested that in order to acquire concepts$_W$ of particular Words presupposes having a concept$_W$ of what Words are, in general. It also seems safe to assume that acquisition of particular concepts$_W$ presupposes possession of the Words that they are concepts$_W$ of. For instance, I assume that one cannot acquire DOG(x) in advance of acquiring the Word 'dog'. For otherwise there would be no experiences of DOG(x)-instances (e.g., 'dog'-utterances, 'dog'-inscriptions, 'dog'-signs) to underwrite the acquisition of DOG(x), or at least not *as such* to invoke Fodor's caveat.[33] That is, I assume that the order of acquisition, in the general case, is that we first acquire Words and then concepts$_W$ of those Words, and perhaps only some time later the concept(s) that those Words lexicalize and are subsequently used to express.

[33] This is not entirely true, as I assume, for example, that a native monolingual speaker of English could acquire a concept of a native Spanish speaker's word 'perro' simply by knowing/being told that 'perro' is a Word of Spanish. The relevant difference is that 'perro' will not be registered as such in the subject's lexicon.

Specifically, I argue in Chapter 4 that Word-acquisition is a graded phenomenon whereby one can acquire a Word, in the sense of having register it as such in one's lexicon, without knowing what that Word means, which is to say without having lexicalized a concept under that Word, and thus without knowing which concept (or concepts) that it can coherently be used to express in ordinary discourse. In such cases, my proposal is that upon acquisition of an otherwise unfamiliar Word (or Word-sound) we use our concept$_W$ of that Word to form an hypothesis *qua* metalinguistic beliefs about what it means. As hypothesized in Chapter 1, this reliance upon concepts$_W$ in the acquisition of lexical meanings is part of what makes metalinguistic-semantic competence (MSC) a crucial ingredient of one's core linguistic-semantic competence (LSC). Before moving on to these details, let me just briefly mention how, on my view, concepts$_W$ become activated as a result of interpreting the meaning of corresponding Words.

3.6. Activating Word-concepts (concepts$_W$)

I have just stated in broad terms how concepts$_W$ are acquired. As with other Fodorian atoms, I assume that once acquired concepts$_W$ are stored somewhere in a speaker's long-term conceptual-semantic memory. That is, by assumption concepts$_W$ are consciously accessible language-*external* mental representations whose representational contents are tacit (i.e., non-conscious, non-conceptual, sub-doxastic) language-*internal* mental representations. I also suggested in Chapters 1 and 2 that Words *lexicalize* the concepts that they are customarily used to express. For instance, the Word 'dog' in the lexicon of most English speakers is used to express DOG(x) *because* speakers have used the former to lexicalize the latter, thereby establishing a fixed semantic relationship

between the two. I also drew my reader's attention to the fact, which I take to be patently obvious, that 'dog' can also be used self-reflexively to express one's concept *of the Word* 'dog', which is to say DOG(w), as in an utterance of the sentence "The word 'dog' contains three letters" or "The word 'dog' designates dogs." As such, it's plausible to suppose that 'dog' lexicalizes both DOG(x) and DOG(w), among perhaps other concepts depending on the speaker's verbal and/or conceptual sophistication.

More generally, the suggestion is that if concept lexicalization is the process by which Words become semantically (i.e., structurally) associated with the concepts they express, and Words can be used self-reflexively to express concepts$_W$ *of themselves*, it follows that concepts$_W$ can become lexicalized under the very Words that they are concepts$_W$ of. In fact, given the close-knit relationship between Words and corresponding concepts$_W$, it would be unsurprising if concepts$_W$ are always the first to be lexicalized under the Words that they are concepts$_W$ of, perhaps as a natural/automatic consequence of having acquired the Word in question. What's more, if Chomsky is right that lexical meanings are understood as instructions to activate the concepts they lexicalize, it follows that concepts$_W$ become activated (or primed/pre-activated) as a natural consequence of interpreting/understanding (the meanings of) the Words that they concepts$_W$ of.[34] I believe this is generally true of all concepts$_W$. However, recall from Chapters 1 and 2 that with respect to proper names such as 'Aristotle', I have argued that the content of the concept$_W$ ARISTOTLE(w) is an integral component of the meaning of 'Aristotle', which we can again think of as expressing the concept$_W$ IS-CALLED(ARISTOTLE(w)). Thus, as a proper constituent of IS-CALLED(ARISTOTLE(w)) the concept$_W$ ARISTOTLE(w) will always

[34] Notice that while Chomsky does not speak of "concept lexicalization" in so many words, I take it that he has something like this mind.

be fully activated simply in virtue of having interpreted the meaning of 'Aristotle'. In Chapter 1 I briefly discussed the potential benefits of this hypothesis wither respect to utterance interpretation, which will become clearer in Chapters 5-7.

3.7. Chapter summary

This concludes my survey of critical background assumptions about the nature of Words, lexical meanings, and concepts that will serve as the core foundation for all that follows. In brief summary, I have proposed that the linguistically-specified, context-invariant meanings of Words are represented by their lexical entries as pointers—what I dubbed CON—that semantically relate those Words to the concepts they lexicalize. And I just argued that the family of expressible concepts will normally include concepts$_W$ of the Words themselves, and which for largely expository purposes I am taking to be Fodorian atoms. With Chomsky, I have in turn hypothesized that lexical meanings are understood by interpreters as instructions to activate one or more of the concepts that their host Words lexicalize—i.e., the concepts that speakers structurally associate with a Word's meaning, again including concepts$_W$ of the Words themselves. And in this way, the meanings of Words guide and constrain without fully determining what they can (and cannot) be used to denote, refer to, or otherwise *talk about* in ordinary discourse.

With the stage now fully set, it is time to move on to the more substantive aspects of my core thesis, the first of which concerns the role of metalinguistic-semantic competence (MSC) in the acquisition of lexical meanings. Subsequent chapters will then address more directly the role of MSC in utterance interpretation.

4. On the Role of MSC in the Acquisition of Lexical Meanings

4.1. Introduction

In this comparatively short chapter I argue that metalinguistic-semantic competence (MSC) is quite often operative in the acquisition of lexical meanings. For in many cases the acquisition of new Words together with their customary meanings does not occur overnight. Rather, developmental studies (not to mention good old-fashioned common sense) indicate that meaning acquisition is a graded phenomenon in that it can take considerable time, conceptual development, and linguistic maturity for Word-learners to gain full control over the meanings of certain Words; see, e.g., Carey (1978), Carey & Bartlett (1978), and more recently Swingley (2010).

With respect to child language development, this appears to be true particularly with respect to the acquisition of verbs and certain functional Words such as prepositions, determiners, etc. In the first instance, Paul Bloom (2000: 25) writes:

> Learning the precise meaning of certain words, especially verbs, might be a long process requiring many trials, as shown by the fact that even some relatively frequent verbs, such as *pour* and *fill*, are not fully understood until middle childhood.

Regarding prepositions, Karmiloff & Karmiloff-Smith (2001: 72) suggest that:

> […] the learning of spatial terms like "in," "on," "under," "in front of," and "behind" is constrained by the child's progressive understanding of the concepts underlying these terms. From this viewpoint, the child would not be able to acquire the meaning of, say, "under" until she had grasped that objects can be positioned one on top of the other.

And even when it comes to common nouns, which children reportedly understand as object names, Eve Clark (2009: 294) comments:

> Some domains take years to acquire, and the meanings children assign to each term may shift as they add words that cut up the conceptual space more finely and learn more about

> how to use each one… In many domains, though, even after several years, children may know little beyond some basic contrasts in meaning. For instance, they may know, by age six, that the words *oak* and *elm* both designate trees, but not be able to identify any instances. That is, they have acquired part of the lexical meaning but they have not yet established the reference for either word. This state of affairs is not unusual: it holds for most adults in some domains as well.

On my view, these remarks suggest that language learners often acquire new Words without a complete grasp of what they mean *conceptually speaking*, which is to say without knowing which concepts other competent speakers use those Words to express. In technical terms, this is to say that competent speakers often acquire Words (or perhaps rather Word-*sounds*) in the absence of an appropriate concept to *lexicalize* under them. Or as Higginbotham (1988) puts it, the child's task is to fill in the "critical elucidations" of the Word's meaning.

The question I explore in this chapter concerns what happens in the interim. My proposal, in brief, is that Word-learners—both young and old alike—use their *concepts*$_W$ *of Words*, or again concepts$_W$, to form *hypotheses qua metalinguistic beliefs* about what otherwise conceptually underspecified Words mean. For while a Word-learner may not possess the concepts that others customarily use certain Words to express, so long as she has acquired the Word in question (i.e., has registered that Word in her mental lexicon) she will be in a position to immediately acquire a metalinguistic concept$_W$ of that Word. She will then be able to use that concept$_W$ to form a corresponding belief/hypothesis about what that Word means. These metalinguistic beliefs then serve as semantic proxies for the non-metalinguistic concepts that learners will eventually acquire, lexicalize, and thus more permanently associate with the Word's meaning. Moreover, I argue that metalinguistic hypotheses about lexical meanings may actually double as the prototypes

that facilitate one's acquisition of lexicalizable concepts. As this relates to my core thesis, if MSC plays a role in the acquisition of lexical meanings, this constitutes an aspect of one's linguistic-semantic competence (LSC), which again to my mind makes MSC a legitimate *explanandum* of semantic theory.

More specifically, I have argued that Words of natural language are type-individuated by their lexically-encoded PHON, SYN, and SEM features, where SEM features are what Chomskyans tend to think of, collectively, as constituting a Word's linguistic meaning. My concern here is specifically with the establishment of what in Chapter 2 I posited to be the 'CON' feature of a Word's lexical entry. Recall that, by hypothesis, CON is a particular SEM feature that semantically relates Words to the concepts they lexicalize. My aim here is to forward a theory about the circumstances under which a speaker finds him/herself in a position to assign values to the CON features of Words thereby fixing a critical aspect of their linguistic meanings. But before launching into the details, it will help to first outline an auxiliary hypothesis about the process of Word acquisition, more generally.

4.2. On the development of metalinguistic competence

To begin, I wish to first briefly consider at what stage of cognitive development children attain metalinguistic awareness, in general, and MSC in particular. Recent studies suggest that the attainment of such awareness/competence might occur much earlier than previously imagined. For instance, it is now well known that by the third trimester prenatal fetuses are highly attuned to speech patterns emanating from outside the womb. Specifically, Guasti (2002: 28) reports that healthy/normal neonates already possess the ability to distinguish utterances of their soon-to-be-acquired native language

from those of unfamiliar foreign languages (or at least those that belong to different rhythmic classes). She adds (p.32):

> Since infants do not know anything specific about their native language (i.e., about phonological constructs, stress, syllables, etc.), the prosody of languages must include some very robust and reliable acoustic cue that infants can easily pick up in a very short time and use for classifying languages.

Furthermore, it has been reported that newborns are highly sensitive to statistical patterns in maternal language that distinguish various lexical categories (e.g., nouns, verbs, etc.) which might later aid in the acquisition of lexical meanings; *cf*. Lany & Saffran (2010). Indeed, a more recent study conducted by Mampe, et. al. (2009) indicates that the crying patterns of newborn babies can be differentiated according to corresponding features their mother's native tongue.

Most relevant for purposes of this chapter, and as briefly mentioned in Chapter 2, there is now ample evidence that infants are born with the capacity to distinguish linguistic utterances as symbols of interpersonal communication—that there is a symbolic link between Words and things in the world that those Words are used to designate. And this is to suggest that linguistic comprehension, and hence linguistic competence, goes beyond mere associative knowledge of the mappings between Word-sounds and their referents. Rather, as Woodward (2004: 149) puts it, "Word learning is both an act of associative learning and an act of symbolic learning." As I am no expert on this topic, let me quote Woodward at length (*ibid.*):

> This aspect of linguistic knowledge rests on more general folk psychological concepts such as attention and intention—when a word is used referentially, the speaker's intention is to draw attention to a particular entity by its use, or to call to mind a particular idea in her interlocutor.

This is not, however, to deny that associative learning plays a role in the child's acquisition of a lexicon. For as Woodward (*ibid*.) adds:

> Both the associative and symbolic aspects of learning are critical. Without the ability to retain and organize associations in memory, it would be impossible to build a lexicon. And, as many theorists have pointed out, the language learning enterprise would not get far without an understanding of the referential nature of the link between words and the world.[1]

Given such evidence, I will take for granted that at least by the time infants begin to understand the meanings of linguistic signals they have acquired a certain degree of metalinguistic-semantic knowledge/competence. Specifically, young language learners appear poised to represent speakers as produced Word-utterances with the intention of using those Words to stand for things in world other than the Word itself. Nonetheless, in order to represent the referential intentions of speakers as such, it would seem that Word-learners must be capable of representing those utterances as Words of natural language.

4.3. On Word acquisition

If what I have said thus far corresponds to an actual language user's semantic psychology, the ability to acquire particular Words presupposes a tacit understanding that Words, in general, are type-individuated by their PHON, SYN, and SEM features. One is therefore given to assume that at some early stage in their linguistic development language learners acquire a universal lexical representation-type, call it √*root*, whose internal structure is the triplet {PHON, SYN, SEM}. That is, while individual Words are themselves unique expression-types, by hypothesis they are each tokens of the same

[1] Woodward directs our attention to Akhtar & Tomasello (2001), Baldwin (1996), Macnamara (1982), and Tomasello (1999) for discussion of this point.

universal type, √*root*.[2] Once acquired, particular Words and their tokens are then individuated by the type-specific values assigned to individual PHON, SYN, and SEM features by the I-languages that generate them. As suggested, however, it often takes time and experience with otherwise conceptually underspecified Words to fully flesh out these values, and in particular the value of their CON feature, which again we might think of as one context-invariant aspect of their lexically-encoded meanings.

So conceived, it seems reasonable to suppose that the first step in acquiring a new Word is to recognize its phonetic form (i.e., acoustic profile) as a potential Word-sound of one's native language, and then to record its role as such in one's lexicon. Imagine, for example, a two-year old learner of English in the throes of acquiring the Word 'apple'. When the child's phonological system first recognizes the sound \'a-pəl\ as a potential Word-sound of English, her I-language registers this event by generating a new lexical entry, i.e., a token of the generic type √*root*, whose initial value assignments can be described by the structure:

(1) √*apple*: {PHON = [\'a-pəl\], SYN = ∅, SEM = {ψ = ∅, CON = ∅}}[3]

I am using 'ψ' as a placeholder for lexically-encoded SEM features other than CON as discussed in Chapter 2 (e.g., [±Count], [±Animate], etc.), and '∅' indicates that the features in question have yet to be assigned values. I will say more about the assignments of SYNs and SEMs presently, and I have already offered a feel for what these features consists in, including major category and subcategorization features. To keep the notation

[2] I am limiting my assumptions here to open-class Words only, thereby allowing for the possibility that closed-class/functional Words have a different structure.

[3] As mentioned in Chapter 2, what counts as a token of the same Word-sound can presumably vary dramatically across dialects and indeed within individual idiolects. Thus, *ex hypothesi* each PHON circumscribes some tolerable *range* of acoustic input/speech signals. This range of signals must then be correlated with the internal feature values assigned to corresponding PHONs by the learner's I-language.

simple, I will represent the phonological component of the Word 'apple', which is to say the lexical root √apple, as PHON(√apple).

In short, one assumes that whatever these PHON features are, by the time a child begins to talk her phonological system knows enough about the sound made by typical utterances of the Word 'apple' to (i) recognize its sound as a potential Word-sound of English, and (ii) encode its role as such in the PHON component of its lexical entry (though the phonological forms of very young language learners are probably at best crude imitations of those of adults). The child's remaining tasks are then to figure out (iii) the grammatical role of 'apple' in her native language, and (iv) its linguistic meaning, which includes identifying which extralinguistic concept (or concepts) that Word being acquired is customarily used to express.[4]

4.4. On the acquisition of conceptually underspecified Words

I also suggested in Chapter 2 that *if* a learner knows (or otherwise has definite beliefs about) which concept an otherwise unfamiliar Word-sound is being used to express, she will be in a position to immediately lexicalize that concept under the Word in question via a process known in the developmental literature as "fast mapping." For instance, if our learner knows/ believes that the Word 'apple' expresses the concept APPLE(x), she will (*ceteris paribus*) lexicalize APPLE(x) under 'apple'. Under current assumptions, this is to say that she will assign the value *@apple* to the CON feature of her lexical entry for 'apple', where *@apple* serves as a pointer to the concept APPLE(x) in her long-term semantic memory. Following notation introduced in Chapters 1 and 2, my

[4] It may be that Word-acquisition for beginning language learners involves a kind of brute association between labels/signals and their meanings, which is to say that complex lexicalization skills of the kind being suggested here may only come online later in the child's linguistic development; one imagines with the acquisition of complex syntax.

hypothesis is that the meaning assignment for the Word ‘apple’ can be characterized as follows:

(2) CON(√apple) = *activate@apple* → APPLE(x)

In terms of comprehension, as described in Chapter 2, we can think of *activate@apple* as being understood by interpreters as an instruction to activate APPLE(x).

While fast mapping helps to explain the vocabulary “explosion” that occurs with children, often beginning around age two, as I indicated above there is also evidence to suggest that Word-learners—both young and old—often acquire new Words on just one or two exposures yet with little or no idea what those Words mean, *conceptually speaking*; that is, without knowing precisely or in some cases even vaguely which *concepts* those Words are customarily used to express. Indeed, learners may have yet to acquire one of the possibly several concepts they will eventually come to associate with the Word’s meaning. In consequence, newly acquired Words are often initially deployed in *conceptually inappropriate* ways.[5] For example, the Word ‘apple’ might be used to in reference to grapefruits, or for that matter any round, edible object.

By around age three, however, children nonetheless begin to deploy conceptually underspecified words in *grammatically appropriate* ways. This ability coincides with the emergence of productive syntax, lending credence to the syntactic bootstrapping hypothesis as exemplified by a child’s tacit understanding of so-called “Jabberwocky” sentences. To borrow an example from Higginbotham (1988),[6] consider (3) below, which has roughly the same *form* of interpretation as (4):

(3) All *mimsy* were the *borogoves*

[5] See Landau & Gleitman (1985), Gleitman (1990), Gillette, Gleitman, Gleitman, & Lederer (1999).
[6] As I understand, this example owes originally to Harman (1974).

(4) The *plates* were all *broken*

While upon first acquaintance a hearer of (3) may lack sufficient evidence to fix the extensions of 'mimsy' and 'borogove', given her tacit knowledge of the grammar, and familiarity with other words in the sentence, by age three or so she can nevertheless infer that 'borogoves' is a plural count noun that designates some class of things, or other. She will also know that 'mimsy' is an ordinary adjective that describes some state or property of 'borogoves' (again whatever those "things" happen to be). As importantly, knowledge of the "logical skeleton" (i.e., grammatical structure) of (3) provides learners with tacit clues about what these words *cannot* mean. Higginbotham (*ibid.*: 166) notes, for instance, that "the child knows that the word 'mimsy' could not mean *tree*, and 'borogove' could not mean *run*." This competence is again adducible to what the child antecedently knows, which is (i) the formal grammatical structure of (3), (ii) the semantic principles by which sentences *with that structure* (or *logical form*) are interpreted, along with (iii) the meanings of its other lexical constituents.[7]

In short, such observations suggest that at some early stage of linguistic development the child's I-language becomes capable of extracting and lexicalizing certain grammatical features of Words (i.e., PHONs, SYNs, and SEMs) based solely on the linguistic context in which they are encountered. Again, these purely formal features of Words often provide learners with tacit clues as to what they can (and cannot) mean, even in the absence of a complete or even vague understanding of which concepts those Words are customarily used to express. Or in my terminology, this kind of tacit lexical

[7] As Lust (2006: 185) explains, "Functional categories provide a critical mechanism by which we map the speech stream to the secret skeleton, just as we do for "Jabberwocky"… Children must link content words to the skeleton… When words are linked to the skeleton, children can label phrases; e.g., nouns will head noun phrases, verbs will head verb phrases.

knowledge is what enables learners to acquire Words and deploy them appropriately prior to having hit upon a stable value to assign to their CON feature.

4.4.1. Innate biases

In addition to their tacit knowledge of syntax/grammar, there is abundant evidence that language learners bring to the acquisition task certain innate *conceptual/semantic* biases that further constrain their initial assumptions about the meanings of newly acquired Words, and particularly with respect to common nouns (which by all accounts dominate the child's early vocabulary). The leading candidates are variously known as the *Whole Object* bias, *Taxonomy*, and *Mutual Exclusivity*, described by Markman (1994: 155-63) as follows:

> ***Whole Object*:** A novel label [i.e., Word-sound] is likely to refer to the whole object and not to its parts, substance, or other properties.
>
> ***Taxonomy*:** Labels refer to objects of the same kind rather than to objects that are thematically related.
>
> ***Mutual Exclusivity*:** Words are mutually exclusive… Each object will have one and only one label.

The *Whole Object* bias is fairly self-explanatory. For instance, when adults use the Word 'dog' in reference to dogs, toddlers naturally assume that the thing designated is the entire animal and not, say, one of its parts, properties, or relations to other things in the environment.[8] According to *Taxonomy*, young children will also naturally assume that 'dog' extends over other instances of the same homogenous class/domain of entities (i.e., that 'dog' is a simple one-place predicate). By contrast, *Mutual Exclusivity*, or what we

[8] Think here Quine (1960), 'gavagai', and his thesis of the "indeterminacy of translation." However, while Quine may have been right that the referents/extensions of words are metaphysically arbitrary, it appears that with respect to language learners they are not conceptually arbitrary.

might call the "no synonym rule," inclines learners to assume, in the absence of evidence to the contrary, that distinct Word-sounds have distinct meanings; i.e., that the relation between Word-forms and their meanings is *one-to-one*. Most theorists also agree, however, that these biases are defeasible in light of negative evidence. Nevertheless, and *ceteris paribus*, they appear to function as important heuristics in the Word acquisition process.

There is in addition a more specific taxonomic principle, the *Basic-Level Category* bias, which naturally favors an interpretation of common nouns as designating a "basic-level" category—a category which, as described by Lust (2006: 285), "has the greatest communicative value." For instance, the Word 'dog' is naturally taken to designate dogs rather than, say, animals (more generally) or collies (more specifically). These biases are not infallible, however, as any parent knows. In particular, young children quite often over-extend (over-generalize) Words like 'dog' to designate any furry four-legged animal, and 'apple' to denote any round edible object (and in one study even a doorknob!); *cf.*, Clark & Clark (1973) and Karmiloff & Karmiloff-Smith (2001).[9] In the other direction, as observed by Lust (2006: 231), young children also under-extend Words such as 'roof' applying it to pitched roofs but not flat ones. Indeed, both over- and under-extensions occur not only with nouns but all open-class lexical items (and, as reported in Lust even classifiers, in classifier languages such as Japanese and Thai).

The key observation, however, is that despite tacitly known constraints on lexical meanings, Word-learners—again both children and adults alike—are quite often not in a position to fix the precise meanings/extensions of Words on first encounter. The obvious

[9] Although, some theorists argue that over-extension occurs because the learner has yet to acquire a more specific Word for the object in question.

question, then, is: What is a learner to do in the meantime? The consensus in the developmental research community is that learners exploit what evidence is available at the time of acquisition to form *an hypothesis* about what the Word means. Now, to put the task in terms of hypothesis formation makes it sound as though the goal of toddlers is to align their lexical meanings with those of adults. Yet it is entirely unclear whether conformity to adult language is ever a concern of young children. For one thing, this presupposes that little kids understand that the meanings they assign to Words might be mistaken. To the contrary, however, it seems that young children quite often simply latch onto the nearest available meaning and just assume that they've got it right, or at least in the absence of negative evidence, which is reportedly quite rare, and perhaps occasionally even in spite of it! But let's set this latter question aside in favor of evaluating what, if anything, learners represent as the meanings of conceptually underspecified Words.

4.4.2. On the role of MSC in the acquisition of lexical meanings

By way of illustration, consider a child who over-extends the Word 'dog' to designate any medium-sized furry four-legged animal, which reflects her belief that 'dog' applies to things other than dogs. More specifically, let us assume that what the child tacitly believes is reflected in (5):

(5) $\exists w$ [DOG(w) & EXPRESSES(w, [FURRY(x) & QUADRUPED(x) & ANIMAL(x)])]

In paraphrase, (5) reflects the child's tacit belief that the Word 'dog' expresses the complex concept [FURRY(x) & QUADRUPED(x) & ANIMAL(x)]. Or differently described, (5) represents the child's working hypothesis about what the Word 'dog' *means*. Indeed, it may be that children initially *lexicalize* something like [FURRY(x) & QUADRUPED(x) &

ANIMAL(x)] under the Word 'dog' *as its meaning*.[10] If that were the case, on my view we could specify its lexically-encoded meaning as (6):

(6) CON(√dog) = *activate@dog* → (FURRY(x) & QUADRUPED(x) & ANIMAL(x))

Of course, by the time a child has acquired FURRY(x), QUADRUPED(x), and ANIMAL(x), he/she is quite likely already in possession of DOG(x). In the imagined scenario, however, the child's predicament is that he/she simply doesn't know (and perhaps doesn't care) that adults use 'dog' to express DOG(x).

Alternatively, it has been suggested to me (by Paul Pietroski, p.c.) that rather than conceiving of the meaning of 'dog' as expressing the complex concept [FURRY(x) & QUADRUPED(x) & ANIMAL(x)], it might be that the child associates the meaning of 'dog' with an extensionally equivalent *atomic* concept, call it BEASTIE(x). The idea is that the child will eventually replace BEASTIE(x) with DOG(x) once she keys in on the fact that others use 'dog' to express DOG(x). If correct, this might obviate the need to generate meaning-hypotheses such as (5). However, it may be nevertheless be the case that learning (5) is a prerequisite for acquiring BEASTIE(x), whereas an even more specific hypothesis will be needed to acquire DOG(x).

To take a more illustrative example, imagine a slightly older child—call her Minnie—who by assumption has the Words 'horse', 'adult', and 'female' in her lexicon, and has correctly lexicalized her concept HORSE(x) under the Word 'horse', her concept ADULT(x) under 'adult', and likewise FEMALE(x) under the Word 'female'. Suppose further that during a visit to the petting zoo, Minnie's father points to a horse and says

[10] I see no reason to deny that some Words lexicalize complex concepts, rather like an idiom or definition/meaning postulate. For example, it would be unsurprising to learn that 'bachelor' lexicalizes [UNMARRIED(x) & MALE(x)]. In the other direction, I see no reason to deny that certain complex expressions (e.g., idioms) lexicalize atomic concepts. For example, it seems to me quite plausible that 'kick-the-bucket' lexicalizes DIE(x).

"Look Minnie, that's a *mare*." Given the principles of *Mutual Exclusivity* and *Taxonomy* from above, Minnie will presumably be inclined to infer, correctly in this instance, that 'mare' is not synonymous with 'horse' but rather designates a subordinate category (i.e., a sub-class of horses). By assumption, however, Minnie knows not which property (or in this case properties) distinguishes mares from the more general class of things called 'horse'. In the absence of further evidence, it may be that Minnie simply introduces an existentially generalized concept to serve as a placeholder for an otherwise unknown property that differentiates mares from the more general category HORSE(x).

Specifically, one might hypothesize that Minnie comes to believe something like (7):

(7) $\exists w\,[\underline{\text{MARE}}(w) \,\&\, \exists F\,[\text{EXPRESSES}(w, [\text{HORSE}(x) \,\&\, F(x)])]]$

In paraphrase, (7) says: For some w, w is the Word 'mare', and for some property *F*, w expresses the complex concept [HORSE(x) & *F*(x)].[11] Here again, it may be that learners initially lexicalize the complex concept [HORSE(x) & *F*(x)] under the Word 'mare' as its meaning. And this is to float the suggestion that [HORSE(x) & *F*(x)] may be what a child who believes (6) actually intends to express with the Word 'mare'. Alternatively, and perhaps more accurately, what the child hypothesizes may be more like (8):

(8) $\exists x\,\exists F\,[F(x) \leftrightarrow \text{HORSE}(x) \,\&\, \exists w\,[\underline{\text{MARE}}(w) \,\&\, \text{SATISFIES}(x, w) \leftrightarrow F(x)]]$

[11] If the agent believes that more than one property distinguishes mares from other horses, then $F(x)$ above is perhaps more accurately represented as some set of concepts $\{F_1(x) \,\&\, F_2(x) \,\&\, \ldots F_n(x)\}$. In the present case, in particular, I am not assuming that Minnie knows immediately that being an adult is part of what distinguishes mares from say fillies. Rather, I assume that the null hypothesis for Word-learners, and children in particular, is that there is just one yet-to-be-discovered property that will adequately distinguish mares from other horses.

In rough paraphrase, (8) can be understood as representing the word-learner's belief that there exists some property *F* such that some entity *x* is an *F* iff *x* is a horse (of some as-of-yet unspecified subcategory), and *x* is a satisfier of the Word 'mare' iff *F*(*x*).

A third alternative is that what the child initially acquires/represents is what in Chapter 3 I called a *C-prototype*—i.e., a Roschian-like prototype whose constituents are Fodorian atomic concepts with weights attached to indicate the relative frequency of their instances among the category in question. For instance, rather than (7) or (8) above one might hypothesize that what the child represents is more like the C-prototype in (9):

(9) $\text{MARE}_P = x$: [(HORSE(x), w_1) & (ADULT(x), w_2) & $\exists F$ (F(x), w_3)]

As also suggested in Chapter 3, it appears that acquiring something like a Roschian prototype constitutes a stage in the acquisition of most lexicalizable concepts. If so, then acquiring/learning (9) may be a necessary precursor to acquiring the primitive concept MARE(x).

In any case, it seems obvious that at some point most English speakers eventually learn that mares are adult female horses. And when this happens, they need only substitute [ADULT(x) & FEMALE(x)] for *F*(x) in (7)-(9) above, as say in (10) below, to conform with the conventional use of 'mare' (if again that happens to be the learner's goal):

(10) ∃w [MARE(w) & EXPRESSES(w, [HORSE(x) & ADULT(x) & FEMALE(x)])]

Indeed, I suspect that 'mare' may be one of those rare Words that actually admits to a definition (e.g., that 'mare' = [HORSE(x) & ADULT(x) & FEMALE(x)]). Or anyway I would be unsurprised if Word-learners often resort to building a representation that is equivalent to a definition in the absence of a more specific, perhaps atomic, concept to lexicalize

under otherwise unfamiliar Words. However, if Fodor is right learners will eventually acquire a *bona fide* atomic concept of mares, which is to say MARE(x). As recall that on Fodor's view one cannot think about mares *as such* without a concept of mares *as such*. And if MARE(x) happens to be the only concept there is of mares as such, then presumably one cannot think about mares as such without having first *acquired* the concept MARE(x).

If that's right, under present assumptions the acquisition of MARE(x) will also normally result in the subject's attainment/formation of the following metalinguistic belief:

(11) ∃w [MARE(w) & EXPRESSES(w, MARE(x)]

Or as represented in the lexicon, we can specify the meaning of the lexical entry of the Word 'mare' as follows:

(12) CON(√mare) = *activate@mare* → MARE(x)

More specifically still, I am inclined to think that MARE(x), as opposed to say DOG(x), is a prime example of what Fodor calls a natural kind concept whose acquisition depends on having an empirically informed *theory* of what mares *are*. In general, I am disposed to think that the acquisition of any concept whose instances cannot be distinguished by their superficial/perceptible properties may occur only in virtue of having first formed a complex definitional belief like (10), which at least for the child serves as a kind of informal scientific theory of what mares are.

There are of course many Words whose scientific and even ordinary meanings remain conceptually underspecified throughout a speaker's adult life. Consider the Word 'elm', for example (or compare the Word 'tabulian' from Chapter 1). As Putnam

famously suggested, one might learn at an early age that elms are a kind of deciduous tree yet grow up not knowing (or again even caring) how to identify their instances as such. In my view, this is again to suggest that otherwise competent speakers will have formed a crude hypothesis (i.e., C-prototype) about what the Word 'elm' means to the effect of (13), or alternatively one of the hypotheses in either (14) or (15):

(13) $\text{ELM}_P = x$: [(TREE(x), w_1) & (DECIDUOUS(x), w_2) & $\exists F$ (F(x), w_3)]

(14) $\exists$x $\exists F$ [F(x) $\leftrightarrow$ TREE(x) & DECIDUOUS(x) & $\exists$w [ELM(w) & SATISFIES(x, w) $\leftrightarrow$ F(x)]]

(15) $\exists$w [ELM(w) & $\exists F$ [EXPRESSES(w, [TREE(x) & DECIDUOUS(x) & F(x)])]]

In either case, the relevant claim is once again that the existentially quantified predicate-concept F(x) is introduced to represent some otherwise unknown property (or set of properties) that is believed to distinguish elms from other tree species.

To be sure, I take it to be an observational fact that most adult English speakers are familiar with elms in name only, or perhaps crude description. Yet so long as one has acquired the Word 'elm', or have minimally registered the Word-sound \'elm\ as a Word of English in one's lexicon, one is thereby in a position to acquire the concept$_W$ ELM(w) and then use that concept$_W$ to form a metalinguistic belief/hypothesis about its meaning, such as one of (13)-(15). Specifically, we can think of (13) serving not only as one's hypothesis about the meaning of 'elm' but also as one's (C-) *prototype of elms*. Moreover, should this prototype grow sufficiently sophisticated, say in consequence of experience with elms, perhaps together with an informal theory of what elms are, it might serve to mediate the acquisition ELM(x). That is, in Fodor's terminology the proposal is that learning (13) might constitute a stage in the process of getting metaphysically locked

to the property of *being an elm*. The acquisition of ELM(x) would then result in the rearrangement of the prototype specified in (13) to that of (16) below:

(16) $\text{ELM}_{P'} = \exists w\ [\underline{\text{ELM}}(w)\ \&\ \text{EXPRESSES}(w, \text{ELM}(x))]$

In turn, having acquired ELM(x) will presumably lead to a revision of the speaker's lexical entry for the Word 'elm' along the lines of (17):

(17) CON(√elm) = *activate@elm* → ELM(x)

But more accurately, for each Word considered in this chapter I assume, as argued for to some extent in Chapters 1-3, that competent speakers will have lexicalized under that Word a metalinguistic concept_W of the Word itself. For instance, to be more precise I should represent the meaning of 'elm' in (17) as (18):

(18) CON(√elm) = *activate@elm* → {$\underline{\text{ELM}}$(w), ELM(x)}

Again, I assume the same holds for 'dog', 'apple', 'mare', and so forth for each Word recorded in a speaker's mental lexicon.

All of this being as it may, I agree with Fodor that the formation of metalinguistic prototypes cannot be the only way to acquire basic concepts. For presumably our non-human, non-linguistic ancestors also know how to pull this trick off, I assume by way of what in Chapter 3 I called *F-prototypes* (i.e., prototypes grounded in the *sensorium*). Rather, to be clear I am merely putting forth an empirical hypothesis about the way *language users* may in fact acquire certain *lexicalizable* concepts. And while I have no empirically testable evidence to prove this, our beliefs about the meanings of Words at least seem to have the right structure and content to serve as prototypes for the concepts that we eventually acquire and then lexicalize under otherwise unfamiliar Words. Thus, what better way to kill two birds with one stone than to use metalinguistic beliefs about

lexical meanings as a way of bootstrapping the acquisition of corresponding lexical concepts? Whether or not I am right about the role of metalinguistic beliefs in the acquisition of lexicalizable pre-linguistic concepts, it strikes me as exceedingly plausible that competent speakers use their concept$_W$ of Words to form beliefs/hypotheses about what those Words mean, and in general that such beliefs are instrumental in the acquisition of lexical meanings.

As any of this relates to linguistic competence, it will be remembered that Chomskyans are committed to the idea that acquiring a language is part-and-parcel of one's linguistic competence. Since the lexicon is an integral component of I-languages, it follows that the capacity to acquire Words along with their linguistic meanings constitutes part of one's core linguistic competence. I have proposed here that the acquisition of lexical meanings may very well depend on using concepts of Words (i.e., concepts$_W$) to form beliefs/hypotheses about what those Words mean, and by the same token what they cannot mean. And this is to suggest that the acquisition of lexical meanings quite often depends on the deployment of a speaker's metalinguistic-semantic competence (MSC).

More generally, as suggested back in Chapter 1 I assume that an adequate theory of linguistic competence should minimally specify (i) what competent speakers must know, or represent, in order to understand the meanings of linguistic expressions, (ii) how their meanings constrain what these expressions can and cannot be used/uttered to talk about in ordinary discourse, (iii) how such knowledge is acquired, and (iv) how this knowledge is deployed in the service of comprehension. In the present chapter I have attempted to demonstrate that in cases where speakers don't know what newly acquired Words mean,

they often recruit their MSC to satisfy one or more of these criteria. If correct, then MSC is an essential ingredient of one's overall semantic competence, and hence one's general linguistic competence. However, I will be arguing in chapters to follow that MSC contributes to speaker's core semantic competence in yet other ways.

4.5. Chapter summary

I have briefly argued here that the acquisition of Words with their lexical meanings is often a graded phenomenon. Specifically, child studies indicate that at some early stage of linguistic development, the child's I-language becomes capable of extracting and lexicalizing certain purely syntactic features of words (i.e., SYNs) based solely on the linguistic context in which they are encountered. Importantly, these features provide learners with tacit clues as to what otherwise unfamiliar Words can (and cannot mean), even in the absence of a complete or even vague understanding of which concepts those Words are customarily used to express. However, the absence of a corresponding concept to lexicalize under a Word appears to be no barrier to the acquisition of the Word itself and its subsequent deployment in ordinary conversation. The absence of a corresponding concept to lexicalize under a Word also seems to be no obstacle to acquiring a concept$_W$ of that Word. Indeed, if what I have said is on the right track belief/hypotheses formed with one's concepts$_W$ often facilitates the acquisition of lexical meanings, perhaps by way of facilitating the acquisition of lexicalizable concepts.

5. On the Role of MSC in the Interpretation of Lexical Polysemy

5.1. Introduction

Under the present hypothesis Words are semantically related to the extralinguistic concepts they lexicalize, and subsequently used to express, by way of a pointer that I have dubbed CON (for "concept," roughly speaking). However, I again assume that CON is just one among perhaps several other SEM features encoded by a Word's lexical entry that collectively constitute its context-invariant linguistic meaning. In terms of comprehension, I am following Chomsky in his conjecture that linguistic meanings are understood by interpreters as instructions to conceptual systems "for thought and action." More specifically, by hypothesis lexical meanings are understood by their interpreters as instructions to activate one or more of the extralinguistic (or pre-linguistic) concepts that their host Words lexicalize.

In this chapter I examine a closely related but more refined conception of linguistic meaning defended by Paul Pietroski (2005, and *forthcoming*) that adds some semantic flesh to the syntactic bones of Chomsky's conjecture. Most relevant to my purposes, Pietroski's semantic architecture offers a theoretically attractive way to think about the representation of *lexical polysemy*—the linguistic phenomenon by which a single Word can be used/uttered to express multiple analytically related concepts, or senses. It will be the burden of this chapter to demonstrate how Word-concepts (concepts$_W$) and corresponding metalinguistic beliefs might fit into this picture and, correlatively, how a speaker's metalinguistic-semantic competence (MSC) contributes to the interpretation (i.e., disambiguation) of polysemous Words relative to a context of utterance.

Before reviewing the relevant details of Pietroski's account, it will be noted that not everyone recognizes polysemy as a legitimate linguistic phenomenon—most notably Jerry Fodor and Ernest Lepore. In particular, Fodor & Lepore (1998) argue instead for what they call a purely "denotational" (as opposed to "decompositional") conception of the lexicon according to which lexical meanings are both *univocal* (i.e., *monosemous*) and *atomistic* (i.e., unstructured/non-decompositional). To foreshadow, it will turn out that my construal lexical polysemy is not necessarily incompatible with Fodor and Lepore's atomistic conception of the lexicon; though it does conflict with their denotational conception of Words. But let me begin by making a case for the *psychological* reality of lexical polysemy, and later I will attempt to show how the conceptual representation of polysemy is related to the semantics of the language in a way that is largely compatible with the constraints imposed by Fodor and Lepore, while also relying heavily on Pietroski's minimalistic semantic architecture.

5.2. A case for the psychological reality of lexical polysemy

I should also qualify immediately that the literature on lexical polysemy is vast and thus allows for only a cursory review of the relevant details here. I will initially approach the question from largely a descriptive perspective by classifying putative forms of lexical polysemy and later consider extant proposals for representing polysemy in the lexicon.

5.2.1. Distinguishing homophony from polysemy

Prima facie most if not all open-class Word-forms are in one way or another semantically ambiguous, or at least with respect to their norm-governed interpretations relative to a context of utterance. More precisely, it strikes me and many other

commentators as patently obvious that in many instances the same Word can be used/uttered to express different concepts on different occasions and/or in different contexts. However, Words appear to be semantically ambiguous in different ways. At a very coarse level of description, this difference is reflected in the widely recognized distinction between lexical *homophony* and lexical *polysemy*.[1]

Lexical homophony (or just *homophony* for short)[2] is characterized by Words that have the same phonology/pronunciation but express *unrelated* meanings. To cite a stock example, the English Word-sound \ˈbaŋk\ can be used to designate either a kind of financial institution or a type of land formation (e.g., a river bank), which are of course semantically unrelated entities. As such, ambiguity theorists tend to agree that the Word-sound \ˈbaŋk\ is *homophonous*, which is to say that there are actually two semantically unrelated Words that as a matter of historical accident happen to have the same pronunciation. In technical terms, this is to suggest that the same phonological form is listed in the lexicon under two separate entries with unrelated meanings.[3] Homophones are traditionally distinguished with numerical indices, as in 'bank$_1$' and 'bank$_2$', thereby marking their status as ontologically distinct Words/lexical entries.[4] In short, the semantic relationship between homophones and their meanings is by definition *one-to-one*.

---

[1] Strictly speaking, many theorists only count homophonous/homonymous Words as ambiguous. However, others see things slightly differently. For example, as Fodor & Lepore (1998: 278, fn.12) write "We take ambiguity to be the generic property of which polysemy and homonymy are species." Many theorists in the psycholinguistics literature also refer to polysemy as a form of lexical ambiguity.

[2] Theorists quite often speak of homophony at the phrasal/sentential level as well. However, for purposes of this chapter we can take 'homophony' to mean lexical homophony.

[3] Some theorists describe homophony as *one* phonological form *linked* to two separate lexical entries.

[4] From a definitional standpoint, *homonyms* are Words that are both *pronounced* and *spelled* alike, whereas *homophones* needn't be spelled alike (e.g., 'tale' and 'tail' are by definition homophones). However, by hypothesis Words as they are represented in the mind have no internal "orthography," *per se*, and thus the distinction as manifest in written language is irrelevant for theoretical purposes. Rather, under present assumptions Words are type-individuated by their phonological and semantic properties. Now, from a computational standpoint distinct <PHON, SEM> will be formally distinct mental symbols, but in virtue of encoding distinct features. If one wishes to think of these symbolic differences in terms of a mental

By contrast, lexical polysemy (or just *polysemy*) arises when what appears to be the same Word is used/uttered to express multiple analytically *related* meanings or senses.[5] On my way of thinking, this is to say that lexical meanings are *conceptually underspecified* in that most if not all open-class Words can be used/uttered to express a *range* of formally distinct yet analytically related concepts. Without begging too many questions, the key observation is that *prima facie* the semantic relationship between polysemous Words and their meanings/senses appears to be *one-to-many*.[6] In Chapter 2 I introduced Chomsky's example of the Word 'book' as an instance of *intra*-category polysemy. Recall that 'book' (or the Word-sound \ˈbu̇k\) can be used as a common noun to designate either a particular physical object ("That book is heavy") or, relatedly, its abstract content ("That book is disturbing"), and in some contexts, as we shall see in a moment, it appears that both senses can be expressed simultaneously. But unlike the two unrelated meanings of 'bank', these two senses are clearly related, analytically or semantically, suggesting that the Word 'book' is in fact polysemous as opposed to homophonous.

A perhaps more dramatic example of lexical polysemy routinely cited in the literature occurs with the Word 'paper' which when used as a common noun is

---

representation's internal "orthography," I suppose that's fine. But one can maintain that these differences supervene on differences in a Word's phonological and semantic properties. In any case, some theorists nonetheless refer to the phenomenon in question as homonymy, but more often the term homophony is used and so that's the one I'll employ for purposes here.

[5] Taylor (2003: 644) writes "A defining feature of polysemy is that the various meanings of a word should be related." To say that some Word senses are etymologically related is neither a necessary nor sufficient condition on polysemy. For while not perceived as being so-related, many homophones reportedly have etymologically related meanings. But as pointed out by Koskela & Murphy (2006: 743) what ultimately matters for the empirical study of lexical ambiguity is how speakers *perceive* the semantic relatedness of meanings.

[6] Or at least in adult lexicons. For if the *Mutual Exclusivity* bias from Chapter 4 is correct, young children generally eschew assigning multiple meanings/senses to the same Word-form. However, it's been reported that children as young as four recognize and process certain forms of lexical polysemy, including the concrete versus abstract readings of the English Word 'book'. See Srinivasan & Snedeker (2011) for discussion.

polysemous, among other ways, as between both count and mass readings. For example, in appropriate contexts ‘paper’ can be used to designate a newspaper, the textual content of a newspaper, the material that it is printed on (mass reading), and even the newspaper’s publisher (e.g., “The paper declared bankruptcy”), among other more distantly related things such as a scholarly journal article and student writing assignment, or for that matter just about anything printed on paper. Moreover, in its plural form, ‘papers’ can be used to designate cigarette wrappers or passport documents, or for that matter just about anything made of paper.[7] The fact that each of these senses is related to, and indeed seemingly derives from, the material substance sense of ‘paper’—what we might call its “core” sense—strongly suggests that the Word ‘paper’ is inherently polysemous.

The standard diagnostic used to distinguish instances of homophony from polysemy is to place candidate Words in variable-binding contexts such as anaphora, sentential ellipsis, and verbal “gapping,” and then gauge the resultant semantic effects. Specifically, the unrelated meanings of homophones normally cannot co-occur in these contexts, or when they do the result is an anomalous (absurd/humorous) effect known as *zeugma*. As a means for comparison, observe first that in unbound, uncoordinated contexts such as (1) one can coherently say things like:

(1) Max drove to the $bank_1$ and Minnie drove to the $bank_2$

where non-coindexation of the two tokens of ‘bank’ again reflects their presumed status as homophones—distinct Words with unrelated meanings. If, however, the second token

[7] While this semantic potential is productive, and perhaps in some cases rule-governed, as mentioned earlier there are limits. For example, the word ‘paper’ cannot be used to denote spinach (among lots of other things). And this constraint on the range of possible meanings, I assume, is due to the context-invariant aspect of the meaning of ‘paper’ encoded by its lexical entry (i.e., the value encoded by its CON feature).

of ‘bank’ is elided (omitted), by common hypothesis the trace (or copy) left behind becomes syntactically coindexed with, and thus referentially bound to, its antecedent, as in (2):

(2) Max drove to the *bank*$_1$ and so did Minnie t_1

The result of coindexation (as determined by the grammar) is that there is no natural reading of (2) according to which Max drove to a financial institution whereas Minnie drove to the river’s edge (or *vice versa*). One can of course *strain* to generate a reading where the trace derives its reference deictically and is thereby understood to mean what (1) means: However, doing so in a bound context such as (2) induces a noticeable zeugmatic effect. Cruse (2000: 31) describes the effect in terms *antagonism*:

> Antagonism between two readings of a word means that they are in competition, like the visual construals of a Necker Cube, and can only be processed one at a time.

In short, the “antagonism” between the two unrelated meanings of ‘bank’ on the strained reading indicates that these meanings are unrelated, and hence that the Word ‘bank’ is homophonous.

By contrast, semantic shifts between the related senses of polysemous Words in bound constructions are typically accommodated more seamlessly, sometimes without noticing that a shift in meaning has occurred. For example, competent speakers easily register the slide between the publisher reading of ‘paper’ in the matrix sentence of (3) below and its objectual reading in the embedded clause:[8]

(3) Max lodged a complaint with the *paper*$_1$ when *it*$_1$ stopped arriving on time

In other words, a natural reading of (3) tolerates equivocation between the publisher sense expressed by ‘paper’ and the shifted material sense of the coindexed anaphor ‘it’;

[8] For other distributional differences between pure homophony and lexical polysemy, see Pustejovsky (1995, 1996), Cruse (1986, 2000), Goddard (2000), and Allwood (2003).

the latter reading being invited by the phrase "stopped arriving on time." Similarly, one can say,

(4) Max read the *book*$_1$ on the table and found *it*$_1$ quite interesting

where here the prepositional phrase "on the table" invites an objectual reading of 'book' in the antecedent whereas 'interesting' in the predicate picks up its abstract sense, here again with no noticeable zeugmatic effect.[9]

I shall be refining this picture moving forward, but a common assumption among ambiguity theorists is that lexical polysemy occupies a logical space of context-sensitive meaning between homophony, on the one side, and monosemy on the other, as depicted in Figure 1 below:[10]

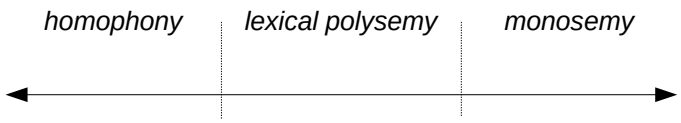

Figure 1: Spectrum of lexical-semantic ambiguity.

5.2.2. On the ubiquity of lexical polysemy

Lexical polysemy is arguably not limited to just common nouns, but rather appears to be ubiquitous across all open-class lexical categories in virtually all studied languages.[11] For example, the English adjective 'heavy' is polysemous as between weighty/massive (as in "a heavy load") versus thick/dense (as in "heavy traffic"). And its

[9] Indeed, as Cruse (2000: 41) observes (and as seconded by Chomsky): "Speakers are not normally aware of the dual nature of *book*. The facets [senses, roughly speaking] form a gestalt. The default usage of *book* is the one which combines the facets."
[10] I wonder if there even are such things as monosemous Words outside of the various and sundry functional vocabularies of natural languages, but we needn't settle this question here.
[11] Several theorists argue that certain functional categories also exhibit lexical polysemy, though I have no need to consider this question here. Rather, my goal is simply to establish that some/many Words are genuinely polysemous.

adverbial form 'heavily' follows the same pattern.[12] Among its myriad other related senses the verb 'walk' is polysemous as between a manner of motion ("Max walked home") and a causative activity ("Max walked the dog"). Now, ambiguity theorists often restrict their analyses to related senses of Words within the same category, or what I referred to in Chapter 2 as *intra*-category polysemy. However, lexical polysemy appears to constitute a class of semantic phenomena that spans categorial boundaries to produce what I labeled *inter*-category or *type*-polysemy.

As hypothesized in Chapter 2, type-polysemy is the result of a morpho-syntactic process known as *type conversion* that generates, for example, de-nominalized verbs ("Max *bottles* wine for a living"), de-nominalized adjectives ("Max only drinks *bottled* water"), de-verbalized nouns ("Max went for *a run*"), de-adjectival adverbs ("The decision weighed *heavily* on Max's mind"), and so forth. Indeed, by all appearances most lexical roots can again be quite easily coerced by the syntax into playing non-canonical grammatical roles. While the senses of Words clearly shift with changes in grammatical type, a characteristic property of inter-category polysemy is that the analytic/semantic relations between senses are typically preserved, which is why I consider the result of type conversion a form of lexical polysemy.

While much more could be said in favor of the distinction between intra- and inter-category polysemy, the relevant claim here is that polysemy can arise as the result of morpho-syntactic manipulations. Next I briefly review a few more commonly recognized varieties of polysemy that can be used to cross-classify instances of both intra- and inter-category polysemy.

[12] These are examples of what I call below *modulated polysemy*—a form of lexical polysemy wherein the sense of a Word depends on, or is modulated by, the meanings/senses of certain other Words with which it combines.

5.2.3. Varieties of polysemy

I have again distinguished intra-category polysemy from inter-category polysemy (or type-polysemy), which while controversial is not without warrant. However, just about all ambiguity theorists agree that the Word 'polysemy' is itself polysemous in that it applies to a range of distinct yet related phenomena, which here again occupy a logical space between homophony and monosemy. While opinions vary greatly about how to divide this space, I will for purposes here decompose lexical polysemy into three major sub-categories: *systematic polysemy*, *non-systematic polysemy*, and *modulated polysemy*. Figure 2 below depicts these sub-categories in relation to one another and with respect to their semantic distance from the two poles:

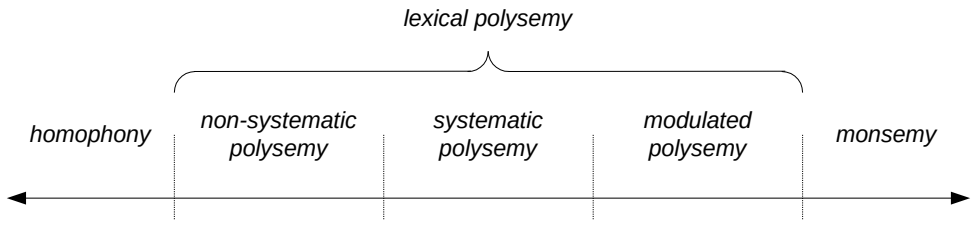

Figure 2: Varieties of lexical polysemy.

Given the proximity of non-systematic polysemy to homophony, their instances are quite often difficult to distinguish. Similarly, the difference between instances of modulated polysemy and monosemy can be quite subtle and therefore difficult to distinguish. Indeed, standard diagnostics designed to distinguish particular cases often generate equivocal results, which leads most researchers to gravitate toward the middle of the spectrum by focusing solely on clear-cut cases of systematic polysemy.

Thus situated, let me briefly elaborate on these distinctions, keeping in mind that the ultimate point of this exercise is to motivate the legitimacy of lexical polysemy as a

bona fide psychological phenomenon that warrants explanation by any complete and respectable semantic theory for natural language. And certain distinctions introduced here will later aid in that effort. It will help to begin in the middle with a brief description of *systematic polysemy*, or what is alternatively known in the literature as "regular," "logical," "metonymical," or "closed-class" polysemy.

The key distinguishing feature of systematic polysemy is that it is *productive* across entire Word sub-classes. This is to say that in regard to instances of systematic polysemy the generation of related senses appears to be *rule-governed* and thus (in most cases) predictable by speakers who know the rule.[13] For example, 'book', 'paper', and 'magazine' are paradigm examples of a general pattern of systematic polysemy between printed objects and their abstract textual contents. A speaker who knows the rule for this pattern knows that it applies to other members of the class (e.g., 'tabloid'). Relatedly, the Word 'newspaper' (or just 'paper') is also part of what's called the "product/producer" paradigm (i.e., newspaper/publisher). Other oft-cited examples include regular alternations between structural openings and their apertures (e.g., 'door', 'window', etc.), containers and their measurements ('teaspoon', 'cup'), sensations/tastes. moods/personalities ('warm', 'cold', 'sweet', 'bitter'), and dance-types versus their performances ('tango', 'waltz'). Such patterns are indeed both extensive and ubiquitous, they occur within all major lexical categories, and again in virtually every studied language (though perhaps not all languages—see Kamei & Wakao [1992] for discussion).[14]

[13] Some theorists, such as Copestake & Briscoe (1995), further distinguish what they call *semi-productive polysemy*, which I won't address here.

[14] In highly inflected languages sense alternations are often accompanied by a change in gender marking. See Soler & Marti (1993) for some examples.

Many researchers contrast systematic polysemy with what is variously known as *non-systematic*, or "irregular," or "metaphorical," or "open-class" polysemy. The key distinction here, as often described, is that instances of non-systematic polysemy are language and/or culture-specific, non-productive, non-rule-governed, and thus generally unpredictable/arbitrary. Instances of non-systematic polysemy must therefore be learned/memorized on a case-by-case basis and then (putatively) recorded as such in a speaker's lexicon.[15] For example, the meaning of 'column' in the noun phrase 'news column' is semantically related to its occurrence in the phrase 'marble column' (i.e., they both involve the vertical arrangement of relevant items). However, extending the meaning of 'column' to the news domain is largely parochial to Anglo-American culture. A general rule of thumb frequently cited in the literature is that whereas systematic polysemy derives from *metonymical* extensions of meaning, instances of non-systematic polysemy are said to derive from *metaphorical* extensions (though I am not entirely convinced by the evidence for this distinction). In any case, what does seem true is that metaphorical sense extensions tend to be more "distant" from the core meaning of the Word from which they derive as compared to metonymical sense-extensions, which is why non-systematic polysemy is situated closer to homophony on the scale in Figure 2 above.

Lastly, let me briefly consider what I have labeled *modulated polysemy*, which in the relevant literature usually goes by the name "vagueness" or "indeterminacy" (which I also think is somewhat misleading but will suppress commentary). Modulated polysemy, as understood here, occurs when the contextually-determined sense of a polysemous

[15] One occasionally finds patterns of non-systematic polysemy that crosses language and/or cultural boundaries, which is generally thought to be purely accidental.

Word depends on the meanings of Words with which it combines (or co-occurs). To cite a few common examples, the Word 'fast' in the phrase 'fast car' seems to mean something different as when combined with 'typist' or, say, 'runner'. Likewise, the sense/meaning of 'healthy' in 'healthy food' is not the same as when it combines with 'appetite' versus say 'complexion'.[16] Similarly, the Word 'heavy' in the phrase 'heavy traffic' does not refer to the weight of the traffic but rather something like its spatial density.

In short, it appears that the various senses of 'fast', 'healthy', and 'heavy' are precisified as determined by the meanings of the nouns they modify. Put differently, the cases just surveyed are all instances of what Copestake & Briscoe (1995: 31) call "adjectival premodification." One notices, however, that Words need not combine directly in order to exhibit modulated polysemy. To borrow an example from Recanati (2004: 34), consider (5):

(5) The city is asleep

As Recanati (*ibid*: 35) observes:

> In one case the speaker means that the inhabitants of the city are sleeping, in the other she means that the city itself is quiet and shows little activity…

More specifically, on Recanati's view if 'The city' is assigned a literal interpretation, then 'asleep' must receive a non-literal (or metaphorical/metonymical) interpretation, and *vice versa*. The point to notice is that in (5) the interaction of senses is between subject and direct object as opposed to a Word and its modifier. In addition, in the latter example the interaction appears to operate in both directions.

[16] To test one's intuitions about this distinction, one can ask whether it is possible to simultaneously assert and deny a sentence like "My appetite is healthy but my diet is not" without contradiction. If the answer is yes, then there is a conceptual difference between the two senses expressed indicating that the Word in question is polysemous.

What is especially interesting about modulated polysemy is that there are infinitely many combinations of Words whose precise meanings/senses cannot be predicted in advance of their appearance in a given context. Moreover, given the open-endedness of modulated polysemy its instances presumably cannot be listed as such in the lexicon (or at least not all of them) but rather must be computed, as it were, on-the-fly. However, competent speakers nevertheless quite easily register the often subtle conceptual differences between unfamiliar Word combinations, suggesting that like systematic polysemy, modulated polysemy is somehow rule-governed.

This latter observation is precisely what motivated James Pustejovsky's (1991, 1995) notion of a *Generative Lexicon* (GL), which is largely designed to account for the unlimited productivity of what I am calling systematic and modulated forms of lexical polysemy. However, it is also Pustejovsky to whom Fodor & Lepore's (1998) objections are directly addressed, as mentioned earlier.[17] In fact, the reason I bother to trudge through these details is in preparation to respond to those objections. But in aid of this discussion, it will help to first consider some standard proposals for how lexical polysemy is represented in the lexicon, including Pustejovsky's GL which is widely considered a viable solution to the problem of lexical polysemy. Having done so, I ultimately agree with Fodor & Lepore that Pustejovsky's account is untenable on both empirical and philosophical grounds. I will attempt to demonstrate, nevertheless, that there is a way of defending the psychological reality of lexical polysemy against Fodor & Lepore's general rejection of polysemy as a linguistic-semantic phenomenon.

[17] Similar objections are raised in Fodor (1998), Cappelen & Lepore (2005), and Lepore & Hawthorne (2011).

5.2.4. Representing polysemy in the lexicon

As suggested earlier, the semantic relationship between polysemous Words and their related senses appears to be one-to-many. And this is again to suggest that a single Word can be used in appropriate contexts to express any of a potentially wide range of semantically related senses/concepts. An important open question concerns how these related senses are encoded in a speaker's mental lexicon, if at all. Within the space of possibilities, we can broadly distinguish what I will call *full specification views* (FSVs) from *semantically underspecified views* (SUVs). According to FSVs the lexical entries of polysemous Words are fully specified for each of their possible senses, which is to say that these senses are learned on a case-by-case basis and then individually recorded in the lexicon. According to SUVs, polysemous Words are underspecified for their range of possible senses, which relative to a context must therefore be dynamically computed on the basis of either lexical-semantic rules or by way of extralinguistic pragmatic/inferential processes, or perhaps some combination thereof (i.e., as a result of interaction between linguistic and pragmatic computational processes).

With respect to FSVs, perhaps the least complex way to represent polysemy is by means of what is called a *sense enumeration lexicon* (SEL) according to which each sense of a polysemous Word is treated as a separate lexical entry (*cf.*, e.g., Kempson [1977]). On this view, polysemes are treated no differently than homophones with respect to their internal representations apart from the fact that their senses happen to be related. The idea, in other words, is that each sense is listed as a distinct Word that happens to have the same phonology as other Words with related senses. However, most lexical semanticists nowadays reject SEL accounts. The worry is that besides being cognitively

uneconomical, SELs fail to capture the fact that the various senses of polysemous Words are semantically related. And capturing such relations is widely considered a core *desideratum* of lexical semantics. SELs also fail to explain the creativity/productivity of lexical polysemy—i.e., the fact that novel Word senses can be generated "on-the-fly" and, moreover, that these novel senses are often transparently understood by interpreters on first encounter.

The leading alternatives to SELs are what I will lump together under the general heading of *single entry views* (SEVs). According to SEVs, the related senses of polysemes are listed collectively under a common lexical root/phonological form. The main advantage of SEVs over SELs is that they explicitly encode the semantic relatedness of semantically related senses. The most straightforward approach here is what I will call the *undifferentiated list view* (ULV) according to which each sense is afforded equal status in its lexical entry—i.e., senses are listed together in the same entry yet without regard to the centrality of one sense over the others.[18] In terms of processing, the idea is that activation of a lexical item (in the course of its interpretation) will cause activation (or priming) of each of its senses. However, the linguistic and/or extralinguistic context will normally favor one interpretation over the others which, when all goes well, will be the sense that its speaker intended to express. The claim, in short, is that both linguistic and extralinguistic factors conspire to ensure that only the context-appropriate

[18] There is substantial psycholinguistic evidence that the activation of a Word also simultaneously activates or primes semantically and phonologically related Words (called "cohorts" in the psycholinguistics literature).

sense rises to salience and is thereby "selected" to participate in compositional processes.[19]

A variation on this theme is what I will call the *core meaning view* (CMV) according to which the related senses of a polysemous Word derive from a single "core," "default," or "literal" meaning—i.e., *the* meaning encoded in a Word's lexical entry. As suggested earlier, for instance, the various senses of the Word 'paper' all appear to derive from its material sense (i.e., being composed of compressed wood pulp, rice, papyrus, etc.). And as suggested in Chapter 2, this core meaning often coincides with the first concept lexicalized under a Word, which in turn often happens to be the sense with the highest frequency of occurrence in the language.

The proposal is that while all known senses of a polysemous Word are individually listed under a single lexical entry, they are differentially weighted/organized such that senses "closer" to the core meaning have lower thresholds of activation, with the core meaning itself being most readily/rapidly activated in otherwise neutral contexts. In contexts favoring a weaker sense (e.g., more semantically "distant" or less frequently used sense), highly activated senses must be actively suppressed/inhibited (or allowed time to decay) in order to facilitate selection of only the context-appropriate sense by compositional processes. The main advantage of CMVs is that they help explain faster response times in the retrieval of high frequency/core senses as confirmed by many psycholinguistics experiments. In addition, the notion of a "core" sense (or meaning) suggests a basis by which regular/systematic sense extensions are computed according to some rule.

[19] When the term 'context' is used here I shall generally mean relevant aspects of both the *linguistic* and *extralinguistic/pragmatic* contexts.

According to one variant of CMV, which falls under the more general heading of what above I called *semantically underspecified views* (SUVs), *only* the core meaning of a polysemous Word is recorded in its lexical entry. In turn, subordinate or derived senses are thought to always be dynamically computed as the context demands. More specifically, if the core meaning of a Word does not "fit" the context, interpreters will attempt to derive the intended sense, either by way of a lexical-semantic rule that governs its possible extensions (*cf*., Nunberg [1979], Caramazza & Grober [1976]), or inferentially through extralinguistic pragmatic processes (*cf*., Sperber & Wilson [1986], Carston [1997], among others). The chief disadvantage of CMVs, whether of the SUV or FSV variety, is that they struggle to explain how, precisely, the process of sense selection operates—i.e., precisely which factors are relevant in the computation of the sense expressed in a given context. Of greatest relevance here is that the absence of a convincing story about sense selection, according to critics, threatens semantic compositionality. CMVs also offer few suggestions as to how irregular/non-systematic senses are generated and/or understood. Yet again it would seem that the interpretation of unpredictable sense-extensions typically poses little problem for competent language users.

Finally, among the more ambitious attempts to respond to the challenges presented by lexical polysemy is again James Pustejovsky's *Generative Lexicon* (GL) (*cf*., Pustejovsky [1991, 1995]).[20] Pustejovsky's overarching goal is to account for the apparent context-sensitivity of lexical meaning, or as he puts it (1995: 42): "how words can take on an infinite number of meanings in novel contexts." He does so by positing richly structured lexical entries that consist in a combination of lexical-semantic features

[20] Copestake & Briscoe (1995) defend a similar account.

(not in the Chomskyan sense of "feature" but more along the lines described in Chapter 3) in combination with a set of lexically-specified rules of semantic composition. To offer a rough sense of what I am talking about, Figure 3 below depicts the internal structure of a typical Pustejovskian lexical entry:

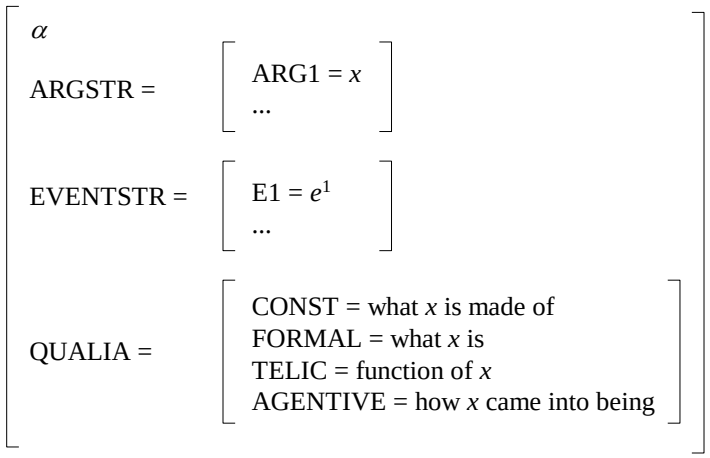

Figure 3: General structure of a Pustejovskian lexical entry.[21]

In Figure 3, 'α' (in the upper left hand corner) represents the Word's ontological category in relation to the hierarchy of worldly entities; e.g., as an *object* versus a *substance*, an *artifact* versus a *natural kind*, etc. 'ARGSTR' represents the Word's *argument structure* (for Words such as verbs, and perhaps prepositions, that have argument structures), 'EVENTSTR' represents its *event structure* (for Words that have event structures), and 'QUALIA', which applies to every open-class lexical item and represents the item's *lexical conceptual structure* (or LCS).

Let me gloss past details pertaining to 'ARGSTR' and 'EVENTSTR' but to say that they encode lexical-compositional *rules* for linking thematic/event participants to their corresponding grammatical argument positions. Most relevant to present purposes is

[21] Figure 3 is reconstructed (though not copied) without permission from Pustejovsky (2005).

Pustejovsky's conception of a Word's *qualia structure*, which as he posits "defines the essential attributes of objects, events, and relations, associated with a lexical item."[22] The attribute (or semantic feature) '[CONST]', for example, represents "the relation between an object and its constituent parts," whereas '[FORMAL]' represents "the basic category of which distinguishes the meaning of a word within a larger domain," '[TELIC]' represents "the purpose or function of the object, if there is one," and '[AGENTIVE]' represents "the factors involved in the object's origins or 'coming into being'."

I will return to some of these details below, but let me emphasize here that Pustejovsky's account is again especially well-suited to accommodate systematic and modulated forms of lexical polysemy. The leading idea being that lexical-semantic information encoded by one Word can interact with that of other Words in a sentence through processes that Pustejovsky (1998: 294) identifies as "co-composition," "type-coercion," and "subselection" (or "selective binding"). Generally speaking, these processes are what determine a Word's contextually appropriate meaning. Importantly, having lexical-semantic rules built into a Word's lexical entry helps to explain how novel Word-senses are dynamically generated. Pustejovsky summarizes his view thusly (2005: 141):

> The richer structure for the lexical entry proposed in GL takes to an extreme the established notions of predicate-argument structure, primitive decomposition and conceptual organization; these can be seen as determining the space of possible interpretations that a word may have. That is, rather than committing to an enumeration of a predetermined number of different word senses, a lexical entry for a word now encodes a range of representative aspects of lexical meaning. For an isolated word, these meaning components simply define the semantic boundaries appropriate to its use. When embedded in the context of other words, however, mutually compatible roles in the

[22] Pustejovsky (2005: 142).

> lexical decompositions of each word become more prominent, thus forcing a specific interpretation of individual words within a specific phrase.[23]

In short, a Pustejovskian GL is somewhat like an SUV in that it does not *list* Word-senses, as such. It is also somewhat like a CMV in that each Word has just one lexical entry and a single lexically-encoded meaning, albeit highly structured in most cases. However, like SEVs Pustejovsky's lexical entries arguably contain enough lexically-encoded semantic information (i.e., semantic attributes/features) to generate not only the various senses of a Word in common currency but also novel extensions to that Word's meaning.

The takeaway from this fairly broad discussion of lexical polysemy is that most theories of the lexicon presume that its lexical entries are complex, structured mental representations that specify or otherwise determine a Word's linguistic meaning, including its possible sense extensions. However, some theorists, and again most notably Fodor and Lepore, find decompositional views of lexical meaning implausible on a number of independent theoretical grounds, both empirical and philosophical, which is ultimately what motivates their rejection of lexical polysemy as a linguistic phenomenon, as I discuss next in greater detail.

5.3. An argument against polysemy

As suggested above, a widely held assumption among linguists and philosophers alike is that any explanatorily adequate lexical-semantic theory must account for facts about the ambiguity and/or context-sensitivity of lexical meanings. Among its other

[23] Pustejovsky (1998: 289) adds: "…I have proposed a framework, Generative Lexicon Theory, that faces the empirically hard problems of how words can have different meanings in different contexts, how new senses can emerge compositionally, and how semantic types predictably map to syntactic forms in language. The theory accomplishes this by means of a semantic typing system encoding generative factors, called *qualia structures*, into each lexical item. Operating over these structures are compositional rules incorporating specific devices for capturing the contextual determination of an expression's meaning."

commonly stated goals, lexical-semantic theory is expected to capture generalizations about the semantic relations between Words such as synonymy, antonymy, mereonymy, hyponymy, and the like (i.e., the so-called "nyms" of traditional lexical-semantic theory). Relatedly, lexical semanticists take themselves to be responsible for explaining materially and/or logically valid inferences licensed by the meanings of Words, as with (6)-(8):

(6) Fido is a dog ⊃ Fido is an animal

(7) Max killed Minnie ⊃ Minnie died

(8) Max is a bachelor ⊃ Max is not married

Furthermore, recall from Chapter 2 that in the wake of Chomsky's *Lexicalist hypothesis,* most generative semanticists nowadays take syntactic structures to be projections of the argument and/or event structures encoded by verbs, which are in turn standardly considered to be part of the verb's linguistic meaning. For instance, Levin & Rappaport-Hovav (2005: 7) note:

> Since the 1980s, many theories of grammar have been built on the assumption that the syntactic realization of arguments—their category type and their grammatical function—is largely predictable from the meaning of their verbs.

As suggested, language theorists generally agree that lexical semantics is responsible, in addition, for identifying the semantic determinants of syntactic structure.

Such a wide diversity of explanatory goals is what drives many semanticists to conclude that lexical meanings must be semantically complex/structured mental representations that decompose into more primitive semantic features, with each feature constituting one facet of a Word's linguistic meaning. Representative of such proposals is again Pustejovsky's *Generative Lexicon* (GL), which is the primary target of Fodor &

Lepore (1998), or just F&L henceforth.[24] Although, the arguments levied by F&L apply more generally to any decompositional conception of the lexicon. As my interest here is specifically with their objection to lexical polysemy as a linguistic-semantic phenomenon, I will limit discussion to what F&L find problematic with Pustejovsky's solution to that question, with the further understanding that F&L's complaints apply, *mutatis mutandis*, to any decompositional theory of lexical meaning. Also for expediency, I shall constrain my review to just one core example. For as F&L state, theirs is a principled objection to the "*kind* of account of the lexicon that [Pustejovsky] endorses," and to which they add "We aren't, in short, just quibbling about cases."

The paradigm cases are instances of what I above described as *modulated polysemy*, where again the precise interpretation of a Word-in-context is seemingly determined in part by the meanings of other Words with which it combines. To borrow one of Pustejovsky's favorite examples, consider the difference in interpretation of the verb 'bake' in the verb phrases 'bake a cake' and 'bake a potato'. In the first case, 'bake' intuitively expresses a *creative process* that requires first assembling some cake ingredients and then heating up the result until it becomes baked; i.e., where the meaning of 'bake a cake' is something like to create by process of baking. That is, cake ingredients do not become a cake until those ingredients are appropriately assembled and sufficiently baked. In the second case, 'bake' seems to behave, semantically, as a mere change-of-state predicate denoting a process that results in a *preexistent* object (e.g., a potato) becoming baked. That is, an unbaked potato is still a potato whereas unbaked cake ingredients is not yet a cake.

[24] *Cf.*, Fodor (2002).

According to Pustejovsky, however, the meaning of 'bake' is univocal in both cases—it expresses the *process of baking*. Rather, the apparent difference in the two senses 'bake', he claims, adduces to differences in the meanings of its respective NP complements in the VPs 'bake a cake' versus 'bake a potato', and in particular to differences in the *qualia structures* of 'cake' and 'potato'. Specifically, 'cake' denotes an artifact kind whereas 'potato' denotes a natural kind. In consequence, the lexical entry for 'cake' specifies a value for the [AGENTIVE] feature of its lexical conceptual structure to the effect that people (agents) create cakes by process of baking.[25] By contrast, the [AGENTIVE] feature of 'potato' is presumably unspecified, as potatoes come into existence through natural processes rather than agentive ones.

The idea, in short, is that the compositionally determined meanings of the VPs 'bake a cake' and 'bake a potato' are determined in part by a lexical-semantic process of "co-composition" where some of the semantic weight falls on the meanings of their complement NPs. Specifically, when 'bake' combines with an artifact term such as 'cake' it generates the creation sense of 'bake', whereas in combination with 'potato' the result is a change-of-state reading. Importantly, however, on Pustejovsky's view it is not the meanings of these constituents that vary with the context but rather the meanings of the *phrases* in which they occur. That is, in each case the meanings of 'bake', 'cake', and 'potato' are each univocal—each Word has a single lexical entry with its decompositional semantic structure fully specified. And this implies that these Words are not themselves *polysemous*. Rather, the idea is that their composed interpretations depend on the meanings of other Words with which they combine.

[25] Pustejovsky (2005: 139).

In reply, F&L observe, first, that its old news that the meanings of complex phrases are compositionally determined by the meanings of their parts. Hence, any difference in the meanings of these phrases is due to a difference in the meanings of one or more of their constituent parts. Yet they contend that from these facts it does not follow that semantic composition occurs in the lexicon, or that the only way to explain the context-sensitivity of lexical meanings is by supposing that lexical entries are semantically complex. Of course, no one denies that the meanings of 'bake a cake' and 'bake a potato' differ at least in part due to differences in the meanings of 'cake' and 'potato', which in these contexts unambiguously denote *cakes* and *potatoes*, respectively. Rather, the question is what explains the perceived difference in the interpretation *of the verb* 'bake' in these phrases? F&L conclude, contrary to Pustejovsky, that this apparent difference in sense is attributable to the fact that the Word 'bake' is itself lexically ambiguous, and more specifically that it is *homophonous*.[26]

F&L's reasoning runs as follows. First, in the negative direction, if Pustejovsky is right that the creative sense of 'bake' is determined by the fact that the meaning of 'cake' specifies that cakes are made by baking, then one presumably does not know the meaning of 'cake' unless one knows how cakes are made. By parallel reasoning, the implication seems to be that if "you don't know how pencils are made, you don't know what *pencil* means [i.e., the Word 'pencil']," which is of course outlandish.[27] On the other hand, if the creation sense of 'bake a cake' is to be explained by the fact that 'cake' denotes an

[26] As I understand him, Fodor takes apparent cases of polysemy to be actual cases of homophony, but with respect to corresponding concepts in Mentalese. By contrast, Lepore seems to think that homophony exists in the lexicon.

[27] Indeed, this argument runs parallel to Fodor objection to Inferential Role Semantics (IRS) according to which one does not know the meaning of a Word (or the content of a concept) unless one knows the inferences that its meaning (or content) license.

artifact then we should expect a creation reading of 'bake' in combination with other artifact terms such as 'trolley' and 'knife'. However, as F&L observe 'bake a trolley' and 'bake a knife' both naturally resist the creation sense of 'bake' in precisely the same way that 'bake a potato' does. In the other direction, one can easily imagine a situation where 'bake a cake' resists the creation sense of 'bake' in favor of mere change-of-state reading; for instance, if the cake in question was purchased pre-assembled, so to speak, and then tossed in the oven.

This is a highly condensed version of their counterargument. But in short, F&L contend that such observations point to the fact that the verb 'bake' in these contexts must itself be lexically ambiguous. Specifically, they write (1998: 280):

> If the creative sense of *bake* is determined by something that it inherits from its direct object—and if *bake* and *cake* are themselves univocal—then *bake a cake* must have only the "creative" reading. But, in fact, *bake a cake* is ambiguous. To be sure, one can make a cake by baking it; but also one can do to a (preexistent) cake just what one does to a (preexistent) potato: put it in the oven and (noncreatively) bake it. Since *bake a cake* is ambiguous and *cake* is univocal, it must be that *bake* is lexically ambiguous (specifically, polysemous) after all, contrary to JP's analysis.

F&L later clarify that the correct analysis of 'bake' is not that it is polysemous but rather again that it is *homophonous*. For while there appears to be a clear sense in which the event denoted by 'bake' in both cases involves a process of baking, F&L ask:

> By what criterion do both kinds of baking count as the same process? What decides that the *bake* in *bake a cake* (hence creating one) denotes the same activity as the *bake* in *bake a potato* (hence heating one)? (Whereas, presumably, *bank* is homonymous because the *bank* in *bank a check* counts as a *different* process from the *bank* in *bank a plane*.).

In other words, the most straightforward explanation of the context-sensitivity of 'bake' is that there are actually two Words (i.e., '$bake_1$' and '$bake_2$') with different meanings.

To better understand their general objection to polysemy, recall first that Fodor is card-carrying conceptual atomist. He is also a *lexical atomist*, which is to say that Fodor takes a very spare view of the lexicon, and indeed of human language more generally. In fact, Fodor has famously argued that natural languages have no semantics, *per se*, and thus that linguistic expressions have no meanings. Rather, on Fodor's view expressions generated by the human language faculty are merely *translated* into semantically interpreted formulae of *Mentalese* (*cf.* Fodor [1975]).[28] Lepore, by contrast, is known to those engaged in the semantics-pragmatics debate as a *Semantic Minimalist*. Specifically, Lepore grants that linguistic expressions have meanings but that they are purely denotational in nature, which is to say unstructured and thus non-decomposable. And for purposes of their collaboration Fodor seems willing to indulge as much. In short, the positive view endorsed collectively by F&L (1998: 270) states that:

> […] the lexical entry for *dog* says that it refers to 'dogs'; the lexical entry for *boil* says that it refers to 'boiling'; and so forth. [F&L use single quotes to indicate meaning/content and italics to designate Words].

The same goes, *mutatis mutandis*, for 'cake' and 'potato'. Specifically, under the assumption that 'bake' is homophonous we can predict that 'bake$_1$' refers to *baking$_1$* (something like *create by heating up*), and 'bake$_2$' refers to *baking$_2$* (simply to *heat up*). In other Words, Lepore's approach to meaning is not only denotational but disquotational.[29]

While I agree with F&L that the two senses of 'bake' are not identical, the chief drawback of F&L's spare conception of the lexicon is that we again lose sight of the fact

[28] A similar view is defended by Hornstein (1986), and one is given think that this is more or less Chomsky's view as well.
[29] However, F&L (1998:270) concede that: "…strictly speaking, lexical *entries* are typically complex. But we claim that they are complex in a way that does not jeopardize either the thesis that lexical *meaning* is atomistic, or the identification of lexical meaning with denotation."

that these senses are in fact analytically/semantically related, as both senses license the following inference:

(9) *x*: baked an $x \supset x$ was heated up

What's more, capturing such facts is one *desideratum* of semantic theory that in my view is worth holding onto. As I argue below, however, one can maintain that all Words are monosemous without analyzing them as homophones. Specifically, my proposal is that the CON feature of a Word serves as pointer to a family of analytically related concepts. Importantly, this pointer is fixed at or around the moment of acquisition, by hypothesis points to just one address in long-term conceptual memory, and is therefore univocal. As suggested in Chapter 2, however, a single lexical-semantic address may be home to multiple analytically related concepts. So understood, there is a tolerably clear sense in which most Words are both monosemous and polysemous. However, technically speaking I take all Words to be monosemous, which is to say that their lexically-encoded meanings are univocal/context-invariant.

In short, if we think of the CON feature of a Word as the context-invariant component of its linguistically-specified meaning, then in this sense I agree with Pustejovsky that apparent cases of polysemy are actually cases of monosemy. As indicated in the beginning of this chapter, my response to F&L relies heavily on the Minimalist semantic architecture developed by Pietroski (2005, and *forthcoming*), which includes a psychologically respectable proposal for how to represent lexical polysemy along the lines of what I have been advocating. My contribution to Pietroski's account will again be to demonstrate the role of a speaker's metalinguistic-semantic competence (MSC) in the representation and interpretation of semantically ambiguous Words

(relative to a context). It will pay to begin here with a broad-stroke sketch of Pietroski's semantic architecture and its central motivation.

5.4. On the nature of semantic composition

Initially inspired by the rediscovery of Frege's logic for invented formal languages, the standing orthodoxy among contemporary semanticists is that the primary mode of semantic composition for natural language is *function application* by way of *argument saturation*; *cf.*, Heim & Kratzer (1997) for a representative perspective. On this approach, the semantic correlates of syntactic constituents are assumed to be either (i) inherently "gappy" *polyadic* predicates (i.e., unsaturated predicates of varying adicities/valences), (ii) the logical constants that saturate these gaps, or (iii) functional elements of one sort or another that bind these constituents together. Such theories are thereby committed to the idea that natural languages make available a certain semantic typology that is needed to account for the various ways that polyadic expressions (expressions of varying adicities or valences) manage to compose, recursively by way of function application, to determine their truth, reference, or satisfaction conditions (relativized to a context).

Other options are available, however. I am again referring specifically to Pietroski's revisionary proposal that the fundamental mode of semantic composition for natural language is not function application but rather *predicate conjunction*. Pietroski's version of Conjunctivism is premised on a growing body of evidence that the combinatorial processes of natural language traffic largely in *monadic* (one-place) predicate-concepts, which couched in terms of traditional typology is roughly equivalent to mental representations of type <*e,t*>, or as Pietroski (*forthcoming*) puts it predicates of type 'M'

(for monadic).[30] Given their type-homogeneity, the semantic correlates of syntactic constituents combine, rather promiscuously, by a single operation of predicate conjunction called M-JOIN, which *ex hypothesi* is triggered by syntactic MERGE. The relevant idea, which is an outgrowth of Chomsky's conjecture, is that the output of linguistic-semantic processing is an LF that serves as a kind of "blueprint" or "instruction" for recursively constructing (or "assembling") a complex monadic concept from a finite set of lexicalized monadic-predicate concepts. Roughly speaking, these linguistically-specified monadic concepts are formal abstractions over the polyadic thoughts expressed by our linguistic utterances.

Importantly, if the processes of semantic composition traffic only (or largely) in monadic concepts, this imposes a formal constraint on concept lexicalization under the assumption that ordinary pre-linguistic animal concepts are polyadic. Thus, central to Pietroski's account is the auxiliary hypothesis that concept lexicalization is a *creative* process which takes lexicalizable polyadic concepts and introduces a *monadic analog* of those concepts by way of *predicate abstraction* (or in the case of singular concepts *argument introduction*) that results in the generation of a corresponding one-place (i.e., monadic) predicate-concept. That is, the product of lexicalization are monadic analogs of the pre-lexical polyadic concepts that we typically use Words to express, the former being what Pietroski sometimes calls *I-concepts*.[31] I-concepts can be thought of as open sentences whose semantic assignments are specifiable in terms of Tarskian satisfaction conditions. Thus, while I-concepts have satisfiers they do not, on Pietroski's view, have Fregean *Bedeutungen* (i.e., referents or truth-conditions).

[30] Pietroski does not countenance a truth-conditional semantics for natural language.
[31] Or as Pietroski sometimes puts it (in p.c.), pre-lexical concepts undergo a specific kind of "reformatting" as a result of lexicalization.

On a global scale, the result of concept lexicalization is the generation of a layer of I-concepts that effectively serves as the central interface between language and thought. On the language side of the border, lexical meanings (i.e., Chomskyan SEMs) are understood by interpreters as instructions to "fetch" basic (formally primitive) I-concepts. In turn, complex LFs—which is to say phrasal/sentential meanings generated by language-internal combinatorial processes—are understood as instructions for conjoining basic I-concepts to produce what we might think of as *I-thoughts*—complex monadic analogs of semantically related polyadic thoughts. The relevant consequence is that for every lexicalized polyadic concept there will be one or more analytically related I-concepts that are structurally associated with a Word's linguistic meaning, more or less as described back in Chapters 1 and 2. And so in this sense the relation between linguistic meanings and the polyadic concepts that speakers lexicalize under Words, and use those Words to express, are mediated by a layer of I-concepts.

5.4.1. Polysemy under Conjunctivism

To help clarify the view, let us consider how one might represent lexical polysemy under Conjunctivism. Take, for example, the English Word 'boil' (or lexical root √boil), which can be used as a transitive/accusative verb to express a two-place predicate of event-*processes*, as in (10) below, *or* intransitively/incohatively as a one-place predicate of event-*states*, as in (11):

(10) Max boiled the water

(11) The water boiled

In other words, (11) describes the terminal *state* of some water having boiled whereas (10) describes Max as the thematic Agent (or Initiator) of a *process*, with some

contextually salient quantity of water as its Theme, that terminated in the water's boiling. (10) and (11) thus provide linguistic-semantic evidence that the same lexical root can be used to express two formally distinct yet analytically related polyadic concepts, let's call them BOIL$_{PROC}$(x, y) and BOIL$_{STATE}$(x). In this case, the content of BOIL$_{PROC}$(x, y) can be described as the process of heating a substance to the point of vaporization, whereas BOIL$_{STATE}$(x) expresses the property of having attained a state of heated vaporization.[32]

Correlatively, the lexicalization of BOIL$_{PROC}$(x, y) and BOIL$_{STATE}$(x) yields two formally distinct yet analytically related *I-concepts*, call them BOIL$_{PROC}$(e) and BOIL$_{STATE}$(e), where 'e' is an event variable introduced by the grammar. Glossing over several important notational details,[33] with respect to comprehension the meaning of 'boil', on Pietroski's view, is understood by interpreters as an instruction to "fetch" one of *either* BOIL$_{PROC}$(e) *or* BOIL$_{STATE}$(e). And this is to say that the *execution* of the meaning-instruction encoded by the lexical entry for the Word 'boil' (or more precisely the lexical root √boil) can be *satisfied* in at least one of two ways.[34] One of course wants to know how this ambiguity is resolved. As Pietroski (2008: 13) puts it, the context-appropriate choice is likely to "reflect complicated *interactions* of grammatical principles with various contingencies of actual language use." I return to this question below. But in most cases, including the present example, the unambiguous linguistic environment is usually

[32] While not identical, notice the similarity here with respect to the meaning of the verb 'bake' from earlier discussion. The present example, however, is better suited to explain not only the 'bake' example but also to highlight a further benefit, as we shall see below.
[33] Pietroski's notation is much more nuanced than represented here. But I am taking certain liberties in the interest of clarity of exposition along with consistency from earlier chapters, I trust without significant loss or gross misrepresentation of Pietroski's broader view.
[34] The concept fetched may be more or less *usable*, conceptually speaking, in a given context as evidenced by Chomsky's oft-cited "Colorless green ideas sleep furiously," which is grammatically impeccable yet conceptually anomalous.

sufficient to settle the matter. Yet as we shall see below the choice is not always straightforward.

As it concerns my project, the relevant aspect of Pietroski's proposal is that for each open-class Word/lexical root—or in my terminology for each {PHON, SYN, SEM} triplet in a speaker's lexicon—there may be a "family" of analytically related I-concepts all clustered around the *same* fetchable address, or "bin" to borrow Pietroski's metaphor, in long-term conceptual semantic memory. While Pietroski does not use this term, I will assume for purposes of this discussion (and for sake of theoretical consistency together with earlier notational conventions) that the address in question is encoded by the CON feature of a Word's lexical entry. More specifically, we can represent the value of the CON feature for the Word 'boil' (i.e., the lexical root √boil) as: CON(√boil) = *fetch@boil*, where *fetch* is a primitive lexical-semantic operation and *@boil* again specifies an address in conceptual memory where a potential family of fetchable monadic I-concepts all "live," so to speak.[35]

If not obvious, the family of fetchable I-concepts lexicalized under a given Word-form can expand (and perhaps contract) over time. Suppose for instance that a Word-learner subsequently acquires the dyadic, pre-linguistic concept BOIL$_{COOK}$(x, y), whose content is something like *to cook by process of boiling*. Like other competent speakers of English, our learner will (again, by hypothesis) lexicalize her newly acquired concept BOIL$_{COOK}$(x, y) under the root √boil thereby adding its monadic analog, BOIL$_{COOK}$(e), to

[35] Notice that Pietroski's notion of "fetching a concept" is roughly analogous to what I have been referring to as "activating a concept." Pietroski's notion of "fetching," however, is tied up with his theory of semantic compositionality and the idea that linguistic meanings are instructions to build/assemble complex monadic concepts. For this reason, Pietroski's notion of "fetching" is much richer than my notion of "activating." With this in mind, and since my proposal is largely neutral about the complexities of semantic compositionality, I will continue to use the term 'activate' with the understanding that 'fetch' is the more appropriate notion.

her idiolect—i.e., to her collection of I-concepts located at the address specified by *@boil*, again each of which she structurally associates with the meaning of the Word 'boil'.

While the semantic relatedness of BOIL$_{COOK}$(e) and BOIL$_{PROC}$(e) is also fairly obvious (they both express predicates of event-processes that involve events of boiling), they have different satisfaction conditions (or as Pietroski sometimes put it, different *application* conditions). For notice that while Max's boiling of the water in (10) entails that some water boiled, his cooking of the eggs by process of boiling as described in (12) below does not entail that the eggs literally boiled:

(12) Max boiled the eggs

Thus, BOIL$_{PROC}$(x, y) and BOIL$_{COOK}$(x, y) are not content-identical concepts, and hence neither are their lexicalized I-concepts BOIL$_{COOK}$(e) and BOIL$_{PROC}$(e). The result of lexicalizing BOIL$_{COOK}$(x, y) under the root √boil, therefore, is the *expanded* family of distinct yet analytically related I-concepts that all cluster around the lexical address specified by *@boil*, and which again following the convention established in Chapters 1 and 2 I will represent as (13),

(13) CON (√boil) = *fetch@boil* → {BOIL$_{PROC}$(e), BOIL$_{STATE}$(e), BOIL$_{COOK}$(e)}

where again the arrow '→' in this context indicates "points to."[36] So construed, we can say, as echoed above and in Chapter 2, that while the meaning of the Word 'boil' (i.e., the lexical root √boil) is *conceptually underspecified* it is nonetheless *semantically determinate*—it is understood by interpreters as *<u>the</u> instruction fetch@boil*, where *@boil*

[36] As alluded to in Chapter 2, it may be that *@boil* points to just *one* of BOIL$_{PROC}$(e), BOIL$_{STATE}$(e) *or* BOIL$_{COOK}$(e)—perhaps the one lexicalized first, or with the highest frequency of use. Either way, nothing much I wish to claim here rests on the outcome. Rather, the present claim depends only on the fact that each of these concepts becomes activated, one way or other, as a result of interpreting the meaning of the Word 'boil'.

points *unambiguously* to a particular address in conceptual memory around which a binful of fetchable I-concepts all reside.[37]

In this way, one can talk about ***the*** *meaning* of the Word 'boil' without specifying precisely which concept that Word is used to express outside of a particular context of utterance. In other words, from a lexical-semantic perspective 'boil' is *monosemous*. Yet there is a sense in which 'boil' is also *polysemous* in that the *concept* fetched in consequence of *executing* its lexically encoded meaning-instruction is context-dependent. And while Pietroski himself is generally skeptical of such evidence (or so far as I can tell), recent psycholinguistic experiments suggest that lexical-semantic interpretation normally results in the initial activation of all semantically related senses of a polyseme. In Pietroski's *terminology*, this suggests that the execution of a meaning-instruction such as *fetch@boil* initially activates *each* I-concept in the bin of fetchable I-concepts specified by *@boil*. However, by hypothesis the operation *fetch* will normally *return* just one I-concept; i.e., *either* BOIL$_{PROC}$(e) *or* BOIL$_{STATE}$(e) *or* BOIL$_{COOK}$(e).[38] We might then think of these analytically related I-concepts as distinct *senses* or *polysemes* of the Word 'boil', yet keeping in mind that 'boil' encodes an unambiguous *linguistic meaning*.[39] Again, precisely which concept is returned as a result of executing a *fetch* operation will normally depend on a number of contextual factors, both linguistic and extralinguistic.

[37] In this regard, Pietroski sometimes uses the metaphor of a single family home or perhaps apartment building with several residents, each of which are regarded by the postal service as living at the same address.

[38] To be clear, Pietroski takes I-concepts to be formally distinct yet syntactically primitive mental representations. My use of subscripting here should not therefore be taken to indicate that I-concepts have hidden internal structure.

[39] If however linguistic meanings are univocal, as suggested, this forces a non-standard definition of the term 'polysemy', yet for good reason.

5.5. Polysemy, Conjunctivism, and Word-concepts

To recap, I have been operating under the hypothesis that Words lexicalize the concepts they are customarily used/uttered to express. Among others, I observed in previous chapters that Words can be used, self-reflexively, to express corresponding *Word*-concepts—i.e., *concepts*$_{W}$ whose representational contents are the very Words (lexical items) used to express them. Given these assumptions, a plausible auxiliary hypothesis is that in addition to the ordinary non-metalinguistic concepts that Words lexicalize, Words also lexicalize corresponding concepts$_{W}$. The chief question I wish to explore here is how my notion of concepts$_{W}$ might fit into the Conjunctivist framework outlined above, and how this extension to the present framework relates to the phenomenon of lexical polysemy.

To a first approximation, one might predict that concepts$_{W}$ are lexicalized under Words in precisely the same way as any other pre-lexical concept such as BOIL$_{STATE}$(x, y), DOG(x), ARISTOTLE$_{PHIL}$, and so forth. For instance, take my concept$_{W}$ of the Word 'dog', which to employ notation introduced in earlier chapters is to say DOG(w). As before, I assume that DOG(w) expresses something like the property of being the English Word 'dog' (as that Word is recorded in my lexicon). Taking Pietroski's proposal at face value, DOG(w) is a lexicalizable concept whose lexicalization should result in the generation of a corresponding monadic analog (I-concept). However, given the intimate relationship between Words and concepts$_{W}$, I am prepared to suggest that concepts$_{W}$ are formally indistinguishable from I-concepts and are therefore lexicalizable in their native format as monadic predicate-concepts. Moreover, as an empirical hypothesis I propose that

concepts$_W$ simply *become* lexicalized, as it were, "automatically" as a natural consequence of having acquired the Words that they are concepts$_W$ of.

As a way of briefly motivating this claim, I have argued that to understand [the meaning of] [40] a Word is to know which concept(s) that it lexicalizes. And to know this much, at least in many cases, seems to entail having first formed beliefs/hypotheses about which concept(s) that Word can coherently be used to express in ordinary discourse (although as suggested in the previous chapter such hypotheses are often mistaken or otherwise incomplete/underspecified). That is, again with Fodor I suppose one could not explicitly think about the meaning of a Word *as such* without first having formed a concept$_W$ of that Word and then having used that concept$_W$ to form a metalinguistic belief/hypothesis about what that Word means. If correct, it strikes me as a plausible working hypothesis that both the acquisition and lexicalization of concepts$_W$ occurs as a natural/automatic consequence of the Word acquisition process. In terms of the natural order of events, it follows that the acquisition and lexicalization of concepts$_W$ normally precedes lexicalization of the non-linguistic (or pre-lexical) concepts that those Words are ordinarily used/uttered to express.

I say "normally precedes" because I wish to leave open the possibility that the acquisition and lexicalization of concepts$_W$ is not a necessary condition for the acquisition and/or lexicalization of *all* pre-linguistic concepts. Rather, it is important to remember that mine is an empirical hypothesis about typical language users in the general case. One might also object that young children lack the kind of explicit metalinguistic knowledge that I have been adverting to and thus, *a fortiori,* lack what I am calling concepts$_W$.

[40] I bracket the expression [the meaning of] because technically speaking it is redundant. For as mentioned back in Chapter 1 I assume that to understand an expression just is to understand its meaning.

However, it has been suggested to me (by Sandeep Prasada, p.c.,) that children may acquire metalinguistic awareness early than previously imagined.[41] For the fact that toddlers do not explicitly talk about language in the way that adults often do is no proof that they lack the target knowledge/competence. However, I am prepared to believe that the acquisition of concepts$_W$, and hence metalinguistic-semantic competence (MSC), emerges at some later stage of development (around age 4-5 by most estimates).[42]

With respect to the present example, my committed proposal is that in addition to other concepts that competent English speakers lexicalize under the Word 'boil' they also lexicalize the concept$_W$ BOIL(w), which by hypothesis is directly fetchable at the address *@boil*. More specifically, while Pietroski himself does not endorse this view I am inclined to think that the concept$_W$ BOIL(w) plays the role of an I-concept in speaker's semantic economy. As again, given the intimate relation between BOIL(w) and 'boil', it strikes me as exceedingly plausible that BOIL(w) becomes lexicalized under 'boil' as a natural consequence of having acquired 'boil'. Given this assumption, it strikes me as uneconomical to think that one's lexicalization of BOIL(w) would result in the introduction of a corresponding I-concept. Rather again, my proposal is that BOIL(w) *just is an I-concept* in Pietroski's sense. If correct, BOIL(w) is directly (i.e., structurally) related to the meaning of the Word 'boil' (as encoded by its CON feature) in the same way as other fetchable I-concepts.[43] So conceived, BOIL(w) would itself be a member of

[41] Yet so far as I know the verdict is still out as to when exactly children begin to acquire introspectively accessible metalinguistic knowledge.
[42] When it comes to adult linguistic competence, I am inclined to think that it may be nomologically impossible to acquire certain uniquely human concepts (e.g., JUSTICE, LOVE, etc.) without first acquiring a certain degree of MSC. Yet, I am not wedded to this claim either.
[43] Indeed, one wonders whether Word-concepts might in fact serve as the "bins" posited by Pietroski, but I won't pursue this thought here.

the set of I-concepts activated, and occasionally fetched, by the semantic interpretation of 'boil', as represented by (14):

(14) CON(√boil) = *fetch@boil* → {**BOIL(w)**, BOIL$_{PROC}$(e), BOIL$_{STATE}$(e), BOIL$_{COOK}$(e)}

I have in fact argued (back in Chapter 1) that this must be the case with proper names whose *meanings*, it will be recalled, are on my view most naturally understood to express the essentially metalinguistic concept$_W$ IS-CALLED(PN(w)), which contains as a constituent a concept$_W$ of the name itself (i.e., PN(w)). I return below to consider the treatment of names under Conjunctivism.

One remaining question is how the process of sense selection operates on an otherwise undifferentiated binful of monadic I-concepts. While Pietroski himself says little about sense selection in terms of processing, his proposal for the representation of lexical polysemy is suggestive of an answer, which I briefly speculate about next.

5.6. On sense selection

5.6.1. Sense selection under Conjunctivism

Given present assumptions, the set of fetchable I-concepts located at the address specified by *@boil* includes BOIL(w), BOIL$_{PROC}$(e), BOIL$_{STATE}$(e), and BOIL$_{COOK}$(e). However, one again wants to know precisely which of these concepts is actually fetched in a given context of utterance—i.e., which concept is returned as the result of *executing* the meaning-instruction *fetch@boil*. To see how the process of sense selection might work, reconsider (15)-(17) from above, repeated below, along with their corresponding

logical forms in (15')-(17'), analyzed in neo-Davidson fashion for illustrational purposes (and ignoring tense):[44]

(15) Max boiled the water

(15') ∃e [Agent(e, Max) & Boil(e) & Theme(e, the water)]

(16) The water boiled

(16') ∃e [Boil(e) & Theme(e, the water)]

(17) Max boiled the eggs

(17') ∃e [Agent(e, Max) & Boil(e) & Theme(e, the eggs)]

In the context of (15)/(15'), speakers will fetch their concept BOIL$_{PROC}$(e) because (i) *ex hypothesi* the thematic structure of (15)/(15') calls for a concept that applies to events with two participants, and (ii) because the understood Theme/Patient of this event must be the sort of thing that can literally boil.[45] By contrast, the absence of a thematic Agent in (16)/(16') indicates a concept of event-states, thereby calling for the selection of BOIL$_{STATE}$(e).

The selection of BOIL$_{COOK}$(e) over BOIL$_{PROC}$(e) in (17)/(17') is a bit trickier because both they and their polyadic counterparts are formally identical concepts. But notice that, on Pietroski's view, the content of BOIL$_{COOK}$(e) is extensionally equivalent to its polyadic counterpart BOIL$_{COOK}$(x, y). In turn, the first argument in BOIL$_{COOK}$(x, y) presumably ranges over things (objects/substances) that can be cooked by process of boiling, yet which as we have seen need not literally boil in the process. And competent speakers of English evidently know this, though presumably not as part of their core linguistic

[44] I take no stance here on the exact format/structure of their LFs. I use the term 'neo-Davidsonian' in the sense that these analyses reflect full separation of Agents and Themes/Patients, unlike Davidson's original (1967) analysis.

[45] Whether or not recognition of this latter fact requires access to extralinguistic/pragmatic knowledge I leave as an open question.

competence but rather as part of what in Chapter 1 I described as their conceptual-semantic competence (CSC). In short, the concept fetched as instructed by the meaning of the Word 'egg', for example, will naturally favor selection of BOIL$_{COOK}$(e) over BOIL$_{PROC}$(e).

In this way, I think Pustejovsky is on to something regarding the way the contextually-salient interpretation of Word can depend on the meanings of other Words in the sentences in which they occur. However, as just suggested I am also inclined to the think, *pace* Pustejovsky, that the rules governing such interactions are not *linguistically* encoded but rather more likely pragmatic/inferential in nature.[46] In any case, similar remarks apply to the contextual selection of the concept$_W$ BOIL(w), as expressed by a sentence such as (18), and whose neo-Davidsonian analysis is given in (18'):

(18) By hypothesis, the word 'boil' fetches BOIL(w)

(18') ∃e [Agent(e, 'boil') & Fetch(e) & Patient(e, BOIL(w))]

Here, one can hypothesize that interaction effects between the meaning of the definite description 'the word' will flag the fact that 'boil' is being used self-reflexively (i.e,, metalinguistically) to express the concept$_W$ BOIL(w), which is then selected for activation (and on Pietroski's model fetched for construction). Or anyway this represents my best guess as to how sense selection might proceed under Pietroski's framework.

5.6.2. Sense selection and metalinguistic-semantic competence (MSC)

Despite often having linguistic cues, there are fully ambiguous contexts where the linguistic context alone is insufficient to determine the speaker's communicative intentions—i.e., wherein it is not be entirely or even vaguely clear which concept the

[46] And so far as I know, there is no reason to deny that pragmatic processes compute the correct results quite quickly, and perhaps also non- consciously.

speaker intends to express with which Words. Consider (19) and (20), for example, where multiple possible readings are equally valid in the absence of relevant extralinguistic contextual support:

(19) Chomsky is difficult to read

(20) Mary cannot bear children

(19) can be understood, for example, as saying either that (i) Chomsky (the person) is mercurial or otherwise difficult to predict, (ii) that the ideas conveyed in Chomsky's written work are obscure, or perhaps more literally (iii) that Chomsky's prose is difficult to parse, or that he simply has poor handwriting. Each of these readings apparently involves distinct yet analytically senses of the Word 'read'. Thus, the Word 'read' is clearly polysemous. By contrast, (20) can be interpreted as stating either that Minnie cannot tolerate the presence of children or that she is reproductively infertile.[47] In short, the relevant question is how, in the absence of clear linguistic cues, might competent listeners settle upon an interpretation of these sentences, even at the risk of being mistaken?

My suggestion is that listeners often rely upon their metalinguistic-semantic competence (MSC), and more specifically on their explicit beliefs regarding the meanings of Words such as 'read' and 'bear'. Take the Word 'read', for example. I have argued that every competent English speaker who has the Word 'read' in his/her lexicon has also acquired and lexicalized a metalinguistic concept$_W$ of that Word—call it READ(w). I have further proposed that such speakers will have used their concept$_W$ READ(w) to form explicit metalinguistic beliefs about what the Word 'read' means, which

[47] Many theorists would argue that 'bear' is homophonous, which is perhaps so. However, I take the Word 'read' to be incontrovertibly polysemous.

constitutes what I have called one's conception of its meaning. More specifically, I have suggested that one's conception of the meaning of a Word is a direct conscious reflection of a one's tacit understanding of which *concepts* are coherently expressible with that Word in which contexts as constrained by its lexically-encoded meaning.

For instance, suppose I believe that the Word 'read' can, in appropriate contexts, be coherently used to express the concepts READ$_{PREDICT}$(x, y), READ$_{COMPREHEND}$(x, y), and READ$_{PARSE}$(x, y), corresponding to the three senses of 'read' mentioned in connection with the ambiguity of (19). By earlier hypothesis, I will have used my concept$_W$ $\underline{\text{READ}}$(w) to form the following beliefs:

(21) ∃w [$\underline{\text{READ}}$(w) & EXPRESSES(w, READ$_{PREDICT}$(x, y))]

(22) ∃w [$\underline{\text{READ}}$(w) & EXPRESSES(w, READ$_{COMPREHEND}$(x, y))]

(23) ∃w [$\underline{\text{READ}}$(w) & EXPRESSES(w, READ$_{PARSE}$(x, y))]

By assumption, I have lexicalized each of the extralinguistic concepts in in (21)-(23) above under the meaning of the Word 'read', which on Pietroski's view implies that I will have introduced three analogous monadic I-concepts; e.g., READ$_{PREDICT}$(e), READ$_{COMPREHEND}$(e), and READ$_{PARSE}$(e). On my view, the list of concepts lexicalized with/under the Word 'read' will also include my concept$_W$ of the Word itself, which is to say $\underline{\text{READ}}$(w).

Thus, we might in turn suppose that the meaning of the Word 'read' in my lexicon is encoded as follows:

(24) CON(√read) = *fetch@read* → {$\underline{\text{READ}}$(w), READ$_{PREDICT}$(e), READ$_{COMPREHEND}$(e), and READ$_{PARSE}$(e)}

Again by hypothesis, I will understand the meaning of the Word 'read' as encoded by (24) as an instruction to activate/fetch (and conjoin) one or more contextually-salient

extralinguistic concepts lexicalized under the memory location specified by *@read*. In light of what was said earlier about the intimate association between Word-meanings and concepts$_W$, I assume that my interpretation of 'read' will typically prime or even activate my concept$_W$ READ(w). Moreover, since READ(w) is a proper constituent of each of (21)-(23), one assumes that activation of READ(w) will naturally lead to the activation of pre-activation of certain metalinguistic beliefs, including possibly those in (21)-(23). And this this way *explicit* thoughts/beliefs about a Word's lexically-encoded meaning can be naturally triggered by one's tacit interpretation of that Word.

Generally speaking, it should be unsurprising that one's interpretation of a Word might naturally trigger a chain of spreading activation, or anyhow increased levels of resting activation, across an entire network of associated beliefs, both metalinguistic and non-metalinguistic. Simultaneously, in contexts such as (19) and (20) listeners will be feverishly generating hypotheses about the speaker's communicative/referential intentions in virtue of having used they Words they did in saying what they said. And when the evidence of speaker-intentions is spare, we of course have no choice but to simply take our best shot.

It is of course difficult if not impossible to know precisely which thoughts/beliefs are triggered by which Words in which contexts. However, one can readily imagine in the case of (19) above that given what competent speakers know about "readings" and "difficult" things, one can immediately rule out an interpretation according to which, for example, reading Chomsky is *physically* demanding. Moreover, if one suspects (for whatever reason) that an utterer of (19) above is referring to Chomsky the person (as opposed to his literature), one can immediately infer that reference is being made to a

psychological state or disposition of Chomsky (the person), and not his written work, which will favor selection of the concept READ$_{PREDICT}$(e)). On the other hand, if one understands the Word 'difficult' as adverting to a mental state of the speaker, one can justifiably infer that READ$_{COMPREHEND}$(e) is the intended concept, and so on for other possibilities.

The upshot is that given contexts such as (19) and (20) wherein speaker-intentions are only vaguely known, or completely unknown, listeners are left to their own devices to assign an interpretation to those particular strings of Words. The choice is not random, however. For there may be, as lexical semanticists often claim, some "core" sense of each open-class lexical item that is selected by default. However, in many contexts the default or what are sometimes called the "literal" senses of each Word in the sentence, such as 'Chomsky', 'difficult', and 'read', do not readily combine to generate a coherent thought. In these situations, my proposal is that listeners rely on their *metalinguistic-semantic competence* (MSC), as guided by beliefs such as (21)-(23), to trigger further beliefs about the different senses of Words such as 'read' and 'difficult' in order to generate epistemically-guided inferences to the best interpretation in light of the evidence available. And while this interpretive strategy may be prone to error, employing one's MSC is often a better bet than simply relying on the first thing that pops into one's mind; e.g., the literal/conventional interpretation of a particular string of Words. In this way, I contend, a speaker's MSC often plays an indispensable role in the process of sense selection, and hence in the process of utterance interpretation.

5.6.3. Sense selection and proper names

Ambiguous uses of proper names are, I believe, even more readily accounted for under present suggestions. As you might recall from Chapters 1 and 2, on my view the meanings of names are naturally understood as containing an essentially metalinguistic component. More specifically, many theorists have argued over the years that the meanings of proper names are fundamentally predicative in nature. For example, Russell (1911, 1918) proposed that the meaning of a name can be analyzed as "the object called 'N'," where 'N' stands for the name in question. To navigate around descriptivist criticisms, Burge (1973) recasts Russell's analysis as "is an entity called 'N." Similarly, Katz's (1994) "pure metalinguistic description" theory of names cashes out the meaning of a name as "the thing which is a bearer of `N'." By contrast, Kent Bach's (1981, 1987 2002) nominal description theory holds that the meanings of proper names are semantically equivalent to the definite description "the bearer of 'N'."

Setting aside their differences, notice that in each case the meanings of names range over a restricted domain of entities (i.e., things so-called) rendering them fundamentally predicative in nature. However, each analysis also makes reference to the name itself rendering them fundamentally metalinguistic in nature as well. While I am again partial to a Burge-style account, it is not a primary goal of this work to provide a fully developed semantic theory for proper names (though I do have a considered view on the matter, as outlined in Chapter 1 and Appendix A of Chapter 2). In any case, the point is that many language theorists agree that names are broadly understood as predicates of things so-called (or so-named). My aim here is to briefly evaluate how a listener's MSC figures into the interpretation of proper names in ambiguous contexts.

In brief review, I have argued that when a syntactically primitive name like 'Aristotle' is used predicatively with its meaning it is naturally understood to express the concept_W IS-CALLED(ARISTOTLE(w)), or as qualified in Chapter 1 perhaps instead the concept_W IS-CALLED(*C*, ARISTOTLE(w)) to account for quantificational uses. I have also argued that when names are used as devices of direct singular reference they combine, syntactically, with a referential index/variable to yield a syntactically complex lexical expression (LE) which contains the name in question as a proper constituent. For example, I take it that the name 'Aristotle' used in reference to Aristotle the philosopher yields an indexed LE of the form '[[$_{\text{NP}}$ [$_{\text{PN}}$ Aristotle]][x_i]]' whose compositionally-determined *semantic value* is specified below in terms of its Tarskian satisfaction conditions (where g is an interpretation function that assigns values to variables relative to a sequence, σ):

(26) $g(x_i)^{\sigma} = \sigma(i)$, $i > 0$

(27) $g([[_{\text{NP}}\ [_{\text{PN}}\ \text{Aristotle}]][x_i]])^{\sigma} = \lambda x$: x satisfies 'Aristotle' iff $x = \sigma(i)$ & $\sigma(i)$ is called 'Aristotle'

According to (27), the semantic value of '[[$_{\text{NP}}$ [$_{\text{PN}}$ Aristotle]][x_i]]' relative to any sequence, σ, of entities in the domain is the i-th object in σ. The category label 'PN' attached to '[$_{\text{PN}}$ Aristotle]' ensures that name tokens apply only to objects in the sequence/domain that are in fact so-called. In this way, when the name 'Aristotle' is interpreted as a proper constituent of an indexed LE it will be understood as referring uniquely and rigidly to a particular 'Aristotle'.[48]

Notice that in order to interpret the indexed LE '[[$_{\text{NP}}$ [$_{\text{PN}}$ Aristotle]][x_i]]' one must first (or perhaps simultaneously) interpret the *bare* (i.e., *non-indexed*) nominal expression

[48] Please refer back to Chapter 2, Appendix A, for further details.

'[$_{NP}$ [$_{PN}$ Aristotle]]'. Yet in its bare form '[$_{NP}$ [$_{PN}$ Aristotle]]' is effectively just the name 'Aristotle', which again by hypothesis is naturally understood to express the concept$_W$ IS-CALLED(ARISTOTLE(w)), or again perhaps IS-CALLED(*C*, ARISTOTLE(w)). By contrast, I have suggested that the indexed LE '[[$_{NP}$ [$_{PN}$ Aristotle]][x_i]]' itself, say as used *in reference to* Aristotle the philosopher, is naturally understood as expressing the "doubly unsaturated" relational concept$_W$ IS-CALLED(*C*, *C*), such that the assignments: *C*=ARISTOTLE$_{PHIL}$ and *C*=ARISTOTLE(w) yield the concept$_W$ IS-CALLED(ARISTOTLE$_{PHIL}$, ARISTOTLE(w)).

As a way of encoding these suggestions, I have further proposed to specify the meaning of 'Aristotle' as follows:

(26) CON(√aristotle) = *activate@aristotle* → {ARISTOTLE(w), IS-CALLED(*C*), IS-CALLED(*C*, *C*), ARISTOTLE$_{PHIL}$, ARISTOTLE$_{TYC}$}

One will notice, however, that this regimentation is not directly compatible with Pietroski's semantic framework. As in consequence of having lexicalized the singular concepts ARISTOTLE$_{PHIL}$ and ARISTOTLE$_{TYC}$ under the name 'Aristotle', by Pietroski's lights speakers will have thereby introduced by abstraction the corresponding *monadic* I-concepts ARISTOTLE$_{PHIL}$(_) and ARISTOTLE$_{TYC}$(_). Although, Pietroski seems to agree that the *name* 'Aristotle' itself, as suggested back in Chapter 1, expresses something like the property of being called 'Aristotle'.[49] Pietroski also has a way of encoding the concepts that I have ascribed to our understanding of names in Conjunctivist terms. However, the benefit of hacking through those details here will have little payoff, as the same basic

[49] As I understand, however, on Pietroski's view, strictly speaking, the meaning of 'Aristotle' is understood as expressing the property of being called by the *sound/noise/PF* of the Word 'Aristotle'. But as mentioned back in Chapter 1, I propose not to quibble over this detail.

points can be made in reference to the semantic axiom in (26). Yet I will proceed under the assumption that (26) can in principle be reformulated in Conjunctivist terms.

The relevant claim, again on my view, is that any use of a proper name, 'PN', is naturally understood to express the fundamentally metalinguistic concept$_{W}$ PN(w). For example, I have argued that when the name 'Aristotle' is used self-reflexively in reference to itself, it will be understood as an instruction to activate (or under Pietroski's model fetch and conjoin) its interpreter's concept$_{W}$ ARISTOTLE(w). When used predicatively, it will be understood as an instruction to first activate ARISTOTLE(w) and one of *either* IS-CALLED(*C*) or IS-CALLED(*C*, *C*), and then "saturate" the open *C* position with the concept$_{W}$ ARISTOTLE(w) to generate one of either IS-CALLED(ARISTOTLE(w)) or IS-CALLED(*C*, ARISTOTLE(w)). And when used *referentially*, the name 'Aristotle' will be understood as an instruction to activate IS-CALLED(*C*, *C*) and doubly saturate such that *C* = ARISTOTLE(w) and *C* = {ARISTOTLE$_{PHIL}$ | ARISTOTLE$_{TYC}$}. Notice that in each use of 'Aristotle' the concept$_{W}$ ARISTOTLE(w) is either directly expressed or appears as proper constituent of the concept$_{W}$ expressed. In the latter two cases, ARISTOTLE(w) will be activated in consequence of having activated the concept$_{W}$ expressed. Hence, as I argued in Chapter 1 even referential uses of names express a *metalinguistic* component of meaning. In turn, understanding the various uses of names always involves recruitment of one's metalinguistic-semantic competence (MSC).

I have argued, in addition, that like any other open-class Word speakers who have the name 'Aristotle' in their lexicon will have used their concept$_{W}$ ARISTOTLE(w) to form explicit metalinguistic *beliefs* about what that name means in terms of which concepts it can coherently be used to express in ordinary discourse. Or put in general terms, I have

argued that one's explicit beliefs about of conceptions of Word-meanings are the direct conscious reflection of one's tacit understanding of the various ways in which those meanings constrain without univocally determining what their host Words can and cannot be used/uttered to denote, refer to, or otherwise talk about in ordinary discourse.

For instance, I take it that typical adult English speakers will have used their concept$_{W}$ ARISTOTLE(w) to form the following metalinguistic beliefs:

(27) ∃w [ARISTOTLE(w) & EXPRESSES(w, ARISTOTLE$_{PHIL}$)]

(28) ∃w [ARISTOTLE(w) & EXPRESSES(w, ARISTOTLE$_{TYC}$)]

Here again, the concept$_{W}$ ARISTOTLE(w) is a proper constituent of the beliefs in (27) and (28) whose activation is normally facilitated by interpreting [the meaning of] 'Aristotle'. In turn, activation of (27) and (28) *ipso facto* involves activation of the singular concepts ARISTOTLE$_{PHIL}$ and ARISTOTLE$_{TYC}$. In this way, merely *mentioning* the name 'Aristotle' will often call to mind thoughts about one or more 'Aristotle's of the interpreter's acquaintance. Indeed, just as with the Word 'read' form above, having ready access to this sort of knowledge can come in handy when trying to disambiguate ambiguous uses of proper names.

With respect to the question of sense selection, consider a context in which one overhears an utterance of (29) in casual conversation:

(29) Athens has undergone significant change since Aristotle

Now, without relevant extralinguistic contextual support a competent listener who is duly knowledgeable of both Aristotle the philosopher and Aristotle Onassis cannot immediately deduce which of these two 'Aristotle's is being referred to by an utterance of (29) without further contextual support. In such contexts, my hypothesis is that

interpreters will naturally understand 'Aristotle' as expressing the fundamentally predicative concept$_{W}$ IS-CALLED(ARISTOTLE(w)), or again that concepts$_{W}$ which competent speakers most naturally associate with the *meaning* of 'Aristotle'. Here again, an interpreter's concept$_{W}$ ARISTOTLE(w) will become activated in virtue of having understood the meaning of 'Aristotle'. Yet given the referential ambiguity, my conjecture is that interpreters will immediately begin sifting through their explicit metalinguistic beliefs about what that name means in search of clues. Specifically, activation of (27) and (28) will cause activation of the singular concepts ARISTOTLE$_{PHIL}$ and ARISTOTLE$_{TYC}$. And this may trigger a massive chain of spreading activation lighting up beliefs about Aristotle the philosopher, Aristotle Onassis, related beliefs about Athens, Jackie O., philosophy, and so on.

Importantly, this sequence of interpretive events may eventually lead to an inference to the best interpretation regarding which individual is being referred to in (29). For instance, suppose our subjects recalls hearing about turmoil among leaders of the Greek federal government over the negative economic ramifications of having hosted the 2004 Summer Olympics. Suppose our subject also recalls learning that during the Onassis era, given his business interests and political clout Arie had lobbied hard against hosting the Olympics in Athens because he predicted that the economic burden would be too great.[50] On this basis, our subject might justifiably conclude that the 'Aristotle' under discussion is in fact Aristotle Onassis and not, say, the ancient philosopher.

Importantly, if what I have conjectured is correct, then the chain of pragmatic reasoning that led to this conclusion was initiated by what our subject understood about the meaning of the *name* 'Aristotle', which *ex hypothesi* is largely metalinguistic in

[50] None of this is factual, so far as I know.

nature. And while our subject's reasoning involved mostly non-metalinguistic knowledge, it is unclear whether that knowledge would be accessible if not by means of activating metalinguistic beliefs such as (27) and (28) above. Moreover, simply recognizing an utterance of (29) as being referentially ambiguous might owe to our subject's metalinguistic-semantic competence. All of this in turn points directly to the fact that MSC often plays an important role in utterance interpretation, sometimes expression interpretation, and in this case the resolution of lexical polysemy.

To take a slightly different kind of example, consider Quine's famous logophoric example in (30):

(30) Giorgione was so-called because of his size

To my knowledge, everyone agrees that the pronoun 'his' in (30) is anaphoric on the name 'Giorgione' and thus coindexed as shown in (30'):

(30') Giorgione$_1$ was so-called because of his$_1$ size

Thus, there is an intuitively clear sense in which the name 'Giorgione' in (30) is *used* referentially, which on my view is to say that its meaning in this context will be naturally understood as expressing the concept$_W$ IS-CALLED(GIORGIONE, $\underline{\text{GIORGIONE}}$(w)). However, (30/30') is also a referentially "opaque" context that prohibits substitution of coreferential names. For instance, while the names 'Giorgione' and 'Barbarelli' corefer, substitution of the latter for the former as in (30) below yields a false statement.

(31) Barbarelli was so-called because of his size

The failure of substitution here evidently owes to the fact that the name 'Giorgione' in Italian literally translates as "Big George." And since Giorgione, the High Renaissance Venetian painter, was reportedly large in stature, the thought expressed (30) is true with respect to Giorgione *as such*, yet false with respect to Barbarelli.

A standard diagnosis of the substitution failure between (30) and (31) is that the adjective phrase (or logophor) ‘so-called’, rather like the adverbial phrase ‘as such’, behaves like an anaphor that makes backward reference to connotative features of the name itself. In (30), the name is ‘Giorgione’ which again connotes “largeness in stature,” whereas the name ‘Barbarelli’ does not (‘Barbarelli was part of Giorgione’s surname). Notice that connotative properties are not the sort of thing one would expect to find among the intrinsic semantic features of proper names in a Chomskyan lexicon. Thus, it seems to me that a correct evaluation of the truth-conditions of (30) could not be grounded in what I called back in Chapter 1 an interpreter’s core linguistic competence, or even his/her core semantic competence (LSC).

Rather, an English speaker’s knowledge that the Italian name ‘Giorgione’ connotes “largeness in stature” is the kind of information that is learned and recorded as part what I am calling one’s metalinguistic-semantic competence (MSC). Thus, to a rough approximation we might suppose that such information is recorded a in metalinguistic belief such as (32):

(32) $\exists w$ [GIORGIONE(w) & CONNOTES(w, LARGE-IN-STATURE(x))]

Given present assumptions, it should be clear how (31) could easily become activated in consequence of having interpreted the meaning of the name ‘Giorgione’. And since such knowledge is instrumental in representing the truth-conditions of (30), one might argue that this aspect of MSC is constitutive of one’s core semantic competence. Something quite similar applies to the interpretation of opacity-inducing dialectical variants of proper names such as ‘Harvard’ and ‘London’. For instance, a speaker might believe that \'harverd\ borders the Charles River but deny that \'havad\ does.

The upshot of this discussion is an empirically-guided refinement of what has been known about the meanings of names for a long time. Specifically, when evaluating what it is that a competent speaker must know in order to understand the meaning of a proper name, Kent Bach (2002: 76) observes that:

> …one's knowledge about particular bearers of particular names does not count as strictly linguistic knowledge. Rather, it is in virtue of one's general knowledge about the category of proper names that one knows of any particular name that when used in a sentence (whether as a complete noun phrase or part of a larger noun phrase) it expresses the property of bearing that name.

And Jerrold Katz (1994: 7) seconds this notion by stating that:

> Speakers who know the sense of 'London' know that literal tokens of that proper noun type denote the contextually definite bearer of the name, but that knowledge does not tell them who or what the bearers are, or even whether there are any tokens of the type. Extra-linguistic information of various sorts is required to know who or what the bearers of the proper noun `London' are and whether an utterance or inscription is a literal occurrence of a token of this type. Given this information, the speaker still requires further extra-linguistic information to know which of the bearers is the referent in the context.

Just so, among other claims made in this chapter it seems to me quite possible that understanding [the meaning of] a proper name is to understand it as an instruction to fetch, among perhaps other concepts, one's concept$_W$ of that name.

5.7. Chapter summary

In this chapter I first argued that lexical polysemy is a psychologically real phenomenon that any complete semantic theory for natural language must account for. In my view, Pietroski's internalist/Conjunctivist semantics, which builds on Chomsky's conception of I-languages, offers an attractive and empirically well-motivated framework within which to better understand and perhaps eventually explain this elusive

phenomenon. The main goal of this chapter, however, was to introduce an independently motivated semantic framework in which to more concretely situate the relevance of a speaker's metalinguistic-semantic competence (MSC) in utterance and/or expression interpretation. In particular, I briefly explored how my notion of Word-concepts (i.e., concepts$_W$) might naturally fit into an internalist-Conjunctivist semantics for natural language. In turn, I have briefly argued for a modest extension to Pietroski's semantic framework which I believe is needed to capture the fundamentally metalinguistic aspects of lexical polysemy, and how a competent speaker's employ MSC in the so-called processes of sense selection. To this end, I have sketched a couple of related applications of MSC to demonstrate the psychological/conceptual resources that interpreter's bring to the task of disambiguating semantically ambiguous lexical constituents, and particularly with respect to our natural interpretation of ambiguous uses of proper names. And while I lack hard empirical support for my proposed extension to Pietroski's account, it is in principle an empirically testable hypothesis. Indeed, I would be delighted to see linguists and cognitive psychologists begin to pay greater attention to the potential role of MSC in utterance/expression interpretation.

From a philosophical standpoint, and more generally speaking, there has to my knowledge been just one or two other serious attempts to evaluate the role of MSC in utterance/expression interpretation, and specifically against the backdrop of a Chomskyan conception of language. One effort in particular owes to a recent (2006) monograph by Robert Fiengo and Robert May entitled *De Lingua Belief* which is the culmination of more than a decade of work on the topic. Unfortunately, I only discovered this work after having devoted more than two years myself thinking about the question. As it turns out,

Fiengo & May's proposal exhibits several affinities with my own view, only theirs is much more thorough and ably stated/defended. In addition, Fiengo & May's account is specifically designed as a linguistic (or what I would call metalinguistic) solution to Frege's puzzle about the informativeness of true identity statements of the form 'α is β', where 'α' and 'β' are coreferential proper names, and correlatively failures of substitution of such names in opaque contexts such as belief reports.

While Fiengo & May's proposal has much to recommend, several commentators have observed that it contains a serious (perhaps fatal) defect. Given its parallels to my account, I devote the next chapter to a careful review of Fiengo & May's account. In Chapter 7 I then develop an alternative approach to resolving these puzzles, based on current assumptions, and which I believe overcomes the shortcomings of *De Lingua Belief* while preserving many of its core intuitions.

6. *De Lingua* Beliefs

6.1. Introduction

I have been arguing that speakers have explicit beliefs about the meanings of the Words they utter, and that such beliefs underwrite an aspect of their core linguistic competence. The present chapter evaluates a closely related proposal by Robert Fiengo and Robert May (F&M hereafter) in a recent monograph entitled *De Lingua Belief* (2006), or *DLB* henceforth. As their title suggests, what I have been calling metalinguistic beliefs F&M refer to as *de lingua* beliefs, which is to say beliefs *about language*. Specifically, F&M agree that competent speakers have explicit beliefs about the meanings of Words—or what F&M refer to as their *semantic Assignments*—that "reflect fundamental aspects of our underlying linguistic competence."

While our respective views again have much in common,[1] they also differ in a few crucial respects. Specifically, F&M's proposal is designed in part as a *linguistic* (as opposed to merely *pragmatic*) solution to Frege's puzzle about (i) the informativeness of identity statements containing type-distinct coreferential names, and correlatively (ii) failures of substitution of coreferential names in propositional attitude reports. Importantly, F&M are committed to the claim that *beliefs of Assignments* are sometimes covertly *expressed* as part of the semantic *content* of certain utterances as reflected in their logical forms. However, F&M's analysis appears to suffer from the same problem that it purports to solve, which, for example, is how it might be that a rational subject and competent speaker could know that *Fa* yet reject *Fb* when '*a*' and '*b*' refer to the same

[1] As a matter of propriety, I should mention that many of my thoughts on the topic were developed prior to, and thus independently of, my acquaintance with *DLB*. My own view has nevertheless benefitted immeasurably from the rigor of F&M's analysis of the relevant issues.

thing. By contrast, I believe the metalinguistic framework developed in previous chapters can be applied as a corrective to F&M's account while preserving many of its otherwise sound intuitions.

I postpone presentation of my own positive account for Chapter 7 where I recast F&M's proposal in largely pragmatic (i.e., non-semantic) terms, as outlined back in Chapter 1. The present chapter is devoted to a detailed analysis of F&M's way of defending the core intuitions underlying their view. Since I again share their core intuitions, my review of several details here will stand in support of my own positive account thus easing the burden of the next chapter. Specifically, I will draw on certain details of F&M's proposal as way to motivate the role of $concepts_W$, and hence a speaker's metalinguistic-semantic competence, in the interpretation of identity statements.

To get the ball rolling, I first review Frege's puzzle about the informativeness of identity statements and relatedly the problem of substitution failure in attitude reports. I then outline Kripke's attempt to neutralize Fregean solutions to the second puzzle. This debate will be old hat for many. But my initial goal is to remind ourselves of Frege and Kripke's respective motivations for thinking that any solution to such puzzles, if one exists, will not be metalinguistic in nature. I then turn to a detailed discussion of F&M's account which attempts to reverse this conclusion, followed by an analysis of why their particular proposal ultimately fails. The next chapter again capitalizes on the salvageable insights of *DLB* to show how such puzzles can be explained by appeal to what I have characterized as a speaker's metalinguistic-semantic competence (MSC).

6.2. Informativeness and substitution

Similar to Frege's *Begriffsschrift*, natural languages tolerate multiple Words that are used to denote the same things, which is perhaps most common with coreferential proper names. This affordance of coreference, however, when coupled with a Millian view of the semantics of names, gives rise to the puzzles mentioned above regarding (i) the informativeness of true identity statements containing formally distinct coreferential names, and relatedly (ii) failures of substitution of coreferential names in referentially opaque contexts such as belief reports.

To see why these cases were puzzling for Frege, first recall J. S. Mill's account of proper names according to which the meaning of a name is nothing more than its referent. Formally, *Millianism* **(M)** about names can be stated as follows:

> **(M)illianism:** The meaning of a proper name is exhausted by its reference. Hence, coreferential names have the same meaning and therefore make the same semantic contribution to the proposition expressed by sentences in which those names appear.

In turn, (M) seems to entail a principle of *Substitutivity* **(S)**:

> **(S)ubstitutivity:** Coreferential names are everywhere substitutable, *salva veritate*, which is to say that their substitution in otherwise structurally identical sentences is truth-preserving.[2]

---

[2] While often implicit in these discussions, I take it that virtually all parties to the debate endorse some version of what I will call *Strict Compositionality* (SC), *Truth Conditionality* (TC), and *Propositionality* (P), which can each be stated roughly as follows:

> **(S)trict (C)ompositionality:** The meaning of a sentence is wholly determined by the meanings of its constituent parts (relativized to a context) together with their mode of semantic composition. Thus, any difference in the meanings of its constituents translates to a difference in the meaning of the sentence.
>
> **(T)ruth (C)onditionality:** The mode of semantic composition for natural language is truth-functional (usually taken to be Fregean "function application"), which is to say that the composed meaning of a sentence, relativized to a context, just is or otherwise fully determines its truth, reference, or satisfaction conditions.
>
> **(P)ropositionality:** Propositions (not sentences) are the bearers of truth, and so the truth-conditions of a sentence (or its utterance relative to a context) determine the proposition that it expresses. Thus,

A stronger version of (S) adds that substitution preserves the cognitive significance of the proposition expressed by such sentences, yielding **(S')**:

> **(S')ubstitutivity:** Coreferential names are everywhere substitutable, *salva veritate* and *salva significatione*.

Opponents of (M), however, maintain that both (S) and (S') are false. And if it can be shown that (S) and (S') are in fact false, then (arguably) so too is (M). Frege's puzzle about the informativeness of identity statements directly challenges the truth of (S'), whereas apparent failures of substitution in opaque or intensional contexts calls into question the soundness of (S).

For example, under the assumption that the names 'Cicero' and 'Tully' refer to the same individual,[3] both (S) and (S') imply that their substitution in the following identity statements should be truth-preserving:

(1) Cicero is Cicero

(2) Tully is Tully

(3) Cicero is Tully

While it is widely agreed that the truth-*values* of (1)-(3) covary, which is compatible with the truth of (S), together (M) and (S/S') jointly imply something much stronger, which is that (1)-(3) have the same meanings that determine the same truth-conditions and thereby express the *same thought/proposition*. Frege of course had reason to doubt this stronger claim.

First, that (3) might differ in meaning from both (1) and (2) is evidenced by the observation that (3) appears to be *informative* in a way that (1) and (2) are not. For

---

any difference in the meaning of a sentence translates to a difference in its truth-conditions, which in turn translates to a difference in the proposition expressed.

[3] As suggested in earlier chapters, I do not believe that names themselves bear reference to things in the world. However, following Fiengo & May I will for purposes of this chapter speak as if they do.

consider Max who by stipulation is a rational agent and competent speaker of English. Let us further assume that Max has the names 'Cicero' and 'Tully' in his idiolect (i.e., lexicon) and uses them appropriately, which is to say that he uses those names with what Scott Soames calls their "usual reference." Let us qualify, however, that Max uses 'Cicero' only in reference to someone he believes was a Roman statesman and 'Tully' exclusively in reference to someone he believes was a Roman poet. And this is to suggest that Max does not know/believe that 'Cicero' and 'Tully' corefer. However, Max presumably knows in virtue of his knowledge of grammar and basic logic that (1) and (2) are trivially true and thus thoroughly *uninformative*. Yet Max can evidently know this much without knowing/believing (or assenting to) (3) *if* he fails to believe that there is just one individual who is in fact so-called.[4]

Now, it obviously won't help to disabuse Max of his mistaken beliefs for a third-party informant to reassert the trivial truth of (1) and (2). By contrast, (3), if uttered persuasively, is putatively capable of causing Max to adjust his (false) belief about how many individuals are the bearers of the names 'Cicero' and 'Tully', thereby leading to the conclusion that Cicero and Tully are in fact one and the same person. For Max to accept the truth of (3) thus constitutes what Frege considered to be a *substantive* extension of his knowledge about the identity of a certain individual. It appears in turn that an identity statement is informative only to the extent that the information it expresses or otherwise conveys is sufficient to cause a rational subject to revise his or her mistaken beliefs about the truth of the identity that it states. The relevant question is this: What is it about (3) as

[4] I am of course using English translations of Cicero's given Latin name(s). 'Cicero' is the standard American English translation, whereas in The King's English he is referred to as 'Tully', or so I have read.

opposed to (1) and (2) that might cause a subject such as Max to revise his mistaken beliefs?

6.2.1. Frege's puzzle about the informativeness of identity statements

This is of course precisely what Frege wanted to know. More specifically, Frege's puzzle about the informativeness of true identity statements amounts to this: If, as entailed by (M) and (S/S'), there are no relevant *semantic* (i.e., *referential*) differences between (1)-(3), then why, or in virtue of what, is (3) informative in a way that (1) and (2) are not? In general terms, Frege wondered why identity statements of the form 'a = b' (in formal languages) have a kind of "cognitive significance" that statements of the form 'a = a' lack.[5, 6] In particular, he wondered whether statements of identity express a *relation*, as their grammatical forms suggests, and if so what is the nature of this relation and what are its *relata*? As a starting point, Frege pretty much took for granted that the logico-mathematic symbol '=' expresses the Liebnizian relation of numerical identity (or self-identity), which is to say the relation that a thing bears to itself and no other. So defined, identity is symmetric, transitive, and reflexive. Such statements should therefore admit free inter-substitution of the signs flanking the identity symbol. Yet one might still wonder, as did Frege, what exactly are the *relata* of such statements?

Frege initially considers just two possibilities: identity statements either express a relation between (i) the *signs* (linguistic expressions) that flank the identity symbol, or (ii) the *objects* (individuals) that those signs designate. Siding with (i), he notes, has the consequence that:

[5] As Frege (1879: 217) puts it, "…sentences of the form a = b often contain very valuable extensions of our knowledge and cannot always be justified in an *a priori* manner."
[6] In terms of natural language, we can take the equality sign "=" as equivalent to the English copular verb "is," or its infinitive "to be."

> What is intended to be said by *a = b* seems to be that the signs or names *`a'* and *'b'* designate the same thing, so that those signs themselves would be under discussion; a relation between them would be asserted.

However, Frege also understood that the assignment of names to their bearers is *arbitrary*, which is to say a matter of linguistic convention. It is then equally arbitrary whether any two names designate the same individual. And so to assert 'a = b' is merely to report a *metalinguistic* fact about language use rather than a *substantive* semantic fact about the relevant object of reference. In this sense, statements of the form 'a = b' are no more informative than statements of the form 'a = a'. Taking option (ii) from above, however, has the equally benign consequence that:

> A relation would thereby be expressed of a thing to itself, and indeed one in which each thing stands to itself but to no other thing.

That is, since everything is self-identical, and any rational agent knows this *a priori*, every statement of identity between a thing and itself is *analytically* true and thus thoroughly *uninformative*.

Such considerations are what ultimately led Frege to the conclusion that proper names have both a *sense* and a *reference*, where the *sense* of name, or in neo-Fregean terms its *meaning*, is an abstract, mind-independent entity that uniquely determines (or presents) its *reference*. So conceived, Frege reasoned that identity statements of the form 'a = b' must express a relation between two different senses or "ways" in which type-distinct coreferential names determine (or present) their common object/referent. As Frege put it, the formal linguistic difference between 'a' and 'b' *corresponds to* or is *connected with* a difference in their "mode of presentation." Thus, Frege concluded that

while the names ‘Cicero’ and ‘Tully’ have the same reference they express different senses (or have different meanings).

Let us assume, for instance, that under the conditions in which he acquired the names ‘Cicero’ and ‘Tully’ that Max associates the former with the sense of ‘Cicero’ as “the Roman statesman” and the latter with the sense of ‘Tully’ as “the Roman poet.” By Frege’s lights, it then becomes plausible to suppose that Max might “grasp” the distinct senses of ‘Cicero’ and ‘Tully’, respectively, without thereby recognizing that they determine the same reference. Hence, to learn the truth of (3) from above (Cicero = Tully) is to learn that ‘Cicero’ and ‘Tully’ express two different ways of presenting the same individual. Since senses are both mind- and language-independent entities, Frege considered such an advance to constitute a *substantive/valuable* extension of knowledge.[7]

In short, the common reference of ‘Cicero’ and ‘Tully’ explains why the truth-*values* of (1)-(3) covary, whereas their difference in sense/meaning accounts for the *informativeness* of (3).[8] And this is because, on Frege’s view, the sense of a proper name makes a semantic contribution to the *thought* expressed by sentences in which it appears. Hence, (1)-(3) differ in meaning because they express different thoughts. Moreover, under the assumption that Max might believe the truth of (1) and (2) while not believing (or positively disbelieving) (3) suggests that senses are truth-relevant in opaque contexts. Furthermore, if as Frege assumed (1) and (2) are analytic/necessary whereas (3) is synthetic/contingent, not only do these sentences express different thoughts, they express

[7] Frege argued that it is possible, at least in principle, that a speaker might associate *different* names with the *same* sense, in which case neo-Fregeans would say that those names are strictly synonymous. This point will be relevant in my discussion of Fiengo & May (2006) below.

[8] Notice that any two *tokens* of the same name-type by definition have the same sense that determines the same reference, which is presumably at least part of the reason why ‘a = a’ is uninformative. By contrast, distinct names that express different senses can uniquely determine the same reference, which presumably accounts, in part, for the informativeness of ‘a = b’. I’ll return to this point in discussion to follow.

different *kinds* of thoughts.[9] The upshot is that given their differences in sense, formally distinct coreferential names are *not* everywhere substitutable, *salva significatione*. As such, Frege's invocation of sense amounts to a refutation of (S').

Similar reasoning applies to apparent failures of (S) in propositional attitude reports. Consider, for example, the simple predicative statements expressed by the sentence-pairs in (4)-(5) and (6)-(7) below:

(4) Cicero was a Roman statesman

(5) Tully was a Roman statesman

(6) Tully was a Roman poet

(7) Cicero was a Roman poet

While in this case (4)-(7) each express contingent (i.e., synthetic) truths, according to (S) *if* (4) is true then so too is (5), and likewise for (6) and (7). What's more, by (M) and (S) it follows that (4)-(7) express just two Russellian propositions whose logical forms (ignoring tense) are roughly as follows:

(8) $\exists x$ [Roman (x) & Statesman(x) & x = Marcus Tullius Cicero]

(9) $\exists x$ [Roman (x) & Poet(x) & x = Marcus Tullius Cicero]

But here again, while one might agree that the sentence-pairs (4)-(5) and (6)-(7), respectively, are true in all the same circumstances, there is reason to doubt that they have the same meanings and/or express the same thoughts (or propositions).

In this case, common intuition has it that if Max does not believe (3) (that Cicero = Tully), he can sincerely assent to (4) and (6) while rationally denying the truth of (5) and (7). Furthermore, based solely on Max's verbal behavior, listeners seem entitled to the following belief attributions:

[9] Although this interpretation of Frege has been disputed. See, e.g., Sluga (1986: 58, 60).

(10) Max believes that Cicero was a Roman statesman

(11) Max does not believe that Tully was a Roman statesman

(12) Max believes that Tully was a Roman poet

(13) Max does not believe that Cicero was a Roman poet

The puzzle now becomes this: On the surface, (10) and (11) seem to attribute to Max belief in both a proposition and its negation, suggesting that he is irrational. However, if Max understands the meanings of 'Cicero' and 'Tully', by (M) and (S) he should be in a position to readily infer that (4) and (5) express the same proposition, as do (6) and (7). In other words, from his belief that (4) and (6) are true Max should be able infer the truth of (5) and (7). Max's denial of (5) and (7) therefore suggests either that the attributions in (11) and (13) are unjustified or, despite robust intuitions to the contrary, that Max is in fact irrational.

Thus, in order to preserve intuitions something seemingly has to give, and there appears to be roughly three options: one might argue (i) that despite appearances, Max does not understand the meanings of 'Cicero' and 'Tully', (ii) that our normal practices of attributing beliefs in these situations are unjustified, or (iii) that (S)ubstitutivity, and hence (M)illianism, are both false. It seems safe to rule out (i) on empirical grounds. For if Max uses the names 'Cicero' and 'Tully' appropriately, which is to say only in reference to the man Marcus Tullius Cicero, one would be hard-pressed to deny that he understands their meanings. This leaves just (ii) and (iii) as live options.

Frege again pins the blame on (iii), and for reasons roughly parallel to the informativeness of identity statements. And this is to say that Max's epistemic condition, and hence his verbal behavior, are explained by the fact that 'Cicero' and 'Tully' differ in meaning or express different senses. With respect to attitude reports, however, Frege's

explanation is a bit more sophisticated involving so-called "reference shifting." Setting this complication aside, Frege concludes that substitution of coreferential names in propositional attitude reports is *not* truth-preserving. And so in this case Frege's invocation of *sense* (arguably) serves as a refutation of (S). And here again, if (S) falls then so does (M). There is still option (ii), however, which is explored by Kripke (1979) in his widely-discussed "defense" of (M)illianism.[10]

6.2.2. Kripke to the rescue

More accurately, rather than being a direct defense of (M), Kripke's (1979) *A Puzzle about Belief* attempts to show that Frege's puzzles militate neither against (M) nor in favor of what he calls a "Frege-Russellian" account of proper names, or what I will abbreviate as **(FR)**:

> **(F)rege-(R)ussell:** The meaning of a proper name is *not* exhausted by its reference. Whatever this extra component of meaning turns out to be (e.g., a Fregean *sense*, or some kind of hidden Russellian descriptive content), it is truth-conditionally relevant in opaque contexts.[11]

Kripke's strategy is to demonstrate that Frege-style puzzles can be generated without appeal to the problematic principle of (S)ubstitutivity. Specifically, Kripke abandons (S) in favor of a principle that he calls *disquotation* and whose truth he takes to be "self-evident." In a slightly modified form, what I will call *Disquotation* **(D)** can be stated thusly:

---

[10] While Kripke does not specifically position his rebuttal as a defense of (M)illianism, many commentators agree that that's what it amounts to.

[11] Recall that on Russell's view names are disguised definite descriptions. If one takes the *sense* of a name to be a definite description, or collection of descriptive properties, that uniquely identifies or otherwise determines its reference, then Russellianism and Fregeanism (with respect to names) effectively (or at least operationally) collapse into what I am calling (FR).

> **(D)isquotation:** If an agent who is both rational and linguistically competent reflectively and sincerely assents to 'p', or positively asserts ‘p’, he or she believes *that p*. Conversely, if a reflective speaker sincerely denies ‘p’, then he/she does *not* believe *that p*.[12, 13]

Kripke not only takes (D) to be self-evident but *suggests* that this principle is implicit in our ordinary belief-ascribing practices. I emphasize *suggests* because while not explicit in Kripke’s discussion, there seems to be a corollary to (D) that deserves to be made explicit in a principle that I will call *Attribution* **(A)**:

> **(A)ttribution:** For every sentence ‘p’ endorsed by a speaker per the conditions of (D)isquotation above, third-party listeners are entitled to *attribute* to that speaker the belief *that p*. Conversely, for every sentence ‘p’ explicitly denied by a speaker per the conditions of (D), third-party listeners may attribute to that speaker disbelief *that p*. The attributions governed by these constraints may in turn be *justifiably reported* using the same sentences that the subject of the report would endorse (or deny), if asked, or “content-preserving” *translations* of those sentences.[14, 15]

Now, Kripke argues that the same sorts of *prima facie* paradoxes generated by (S) can also be generated by (D) in conjunction with what I am calling (A). The consequence is this: If such apparent paradoxes falsify (S) then they also falsify (D). However, if one agrees with Kripke that (D) is axiomatic then the problem must lie elsewhere. In short,

---

[12] Kripke actually develops stronger and weaker versions of *disquotation* which I have consolidated here for expediency.

[13] Following Kripke, I assume that 'p' can be replaced, inside and outside quotation marks, by any grammatical sentence of English (unless otherwise qualified). We may also assume for the sake of argument that the sentence replacing `p′ is free from indexicals, pronouns, and other inherently ambiguous or otherwise context-sensitive expressions.

[14] Kripke invokes another principle that he calls *translation*, or what I will describe as *Translation* **(T)**:

> **(T)ranslation:** If a sentence of one language expresses a truth in that language, then any translation of it into any other language also expresses a truth (in that other language).

(T) is specifically employed in Kripke’s so-called “Pierre puzzle.” Given that this puzzle is in my view weaker than his “Paderewski puzzle,” I won’t discuss it here.

[15] The way Kripke frames (D) makes it seem as though the principle is known only to the believer, or the theorist, or perhaps only from God’s point-of-view. (A) is intended to make clear that *listeners* are entitled to these inferences. And while somewhat of a nit, I assume that one can attribute beliefs to speakers —i.e., take something like a pro-attitude towards the speaker’s beliefs—without explicitly *reporting* these attributions, hence the qualification in (A) about reporting. As importantly, if I understand correctly (A) is really the principle responsible for generating the puzzles, as will be made clear in a moment.

Kripke's conclusion is that the puzzles in question arise not from failures of (M), (S), *or* (D), *per se*, but rather from a breakdown in our practices of *attributing* beliefs to other speakers on the basis of what they say. As I interpret him, in other words, our intuitions about failures of (S) can be adduced to failures of (A).[16] Ultimately, Kripke is skeptical about solving the puzzle. The upshot, however, is that his way of *generating* them arguably neutralizes Frege's *reductio* of (S), and hence (M).

Kripke's rebuttal to (FR) is most forcefully demonstrated by his so-called "Paderewski puzzle," which proceeds with the introduction of Peter who, like Max from above, is presumed to be a rational agent and competent speaker. In this case, Peter comes to believe that there are two individuals with the same name, *viz*., 'Paderewski'; one of whom is believed to be an influential politician and the other a talented musician. Unbeknownst to Peter, however, there is in fact just one person so-named who is both a politician and a musician. In addition, Peter is naturally prone to doubt that politicians have musical talent and that musicians have political clout. Given Peter's epistemic state, he sincerely assents to (14) and (15) below in reference to Paderewski "the politician" and likewise to (16) and (17) in reference to Paderewski "the musician:"

(14) Paderewski has political clout

(15) Paderewski lacks musical talent

(16) Paderewski has musical talent

(17) Paderewski lacks political clout

Correlatively, with earlier intuitions now codified in (D) and (A), duly informed listeners appear entitled to the following pair of *prima facie* contradictory belief attributions:

[16] Indeed, Kripke's article might have been better titled *A Puzzle About Belief Attributions*. For Kripke says very little about what he takes to be the underlying nature of *belief* itself.

(18) Peter believes that Paderewski has political clout

(19) Peter does not believe that Paderewski has political clout

(20) Peter believes that Paderewski has musical talent

(21) Peter does not believe that Paderewski has musical talent

The puzzle in this case is over the surprising difficulty in answering the rather straightforward question: *Does Peter, or does he not, believe that Paderewski has musical talent* (or *political clout*)? Kripke's conclusion is that there is no correct answer to this question. For on the assumption that all parties are using the name 'Paderewski' in the "normal" way, which is again to say with its "usual" reference, either Peter has inconsistent beliefs or the attributions of his beliefs are unjustified. And this is to suggest that (D) and/or (A) is false. However, intuitions again push hard in the opposite direction—that neither consequence follows. For as the situation has been described, both Peter and his auditors seem perfectly justified in their verbal behavior.

Notice further that there appears to be no grounds for invoking (FR)-style objections. For under present assumptions (22) appears to have the form 'a = a' (or 'α is α') and is therefore uninformative:

(22) Paderewski is Paderewski

As such, there is no obvious formal difference between names (or their senses) to block what are presumably instances of the trivially valid inference scheme:

(23) *S* believes that $Fa \supset$ *S* believes that *Fa*

Indeed, as Kripke suggests no amount of logical acumen could cause Peter to reverse his verbal commitments apart from acceptance of some further premise that would compel him to believe that there exists just one individual named 'Paderewski'.

As we shall see below, Fiengo and May (2006) argue convincingly that an utterance of (22) in the right circumstances *can* effect such a change in Peter's beliefs, which calls into question the assumption that (22) necessarily expresses a proposition of the form a = a. But for now, Kripke's point is that without a formal difference between names to point to, the question of substitution failure does not arise. For the "Paderewski" puzzle can be generated without invoking the problematic principle of (S)ubstitutivity.

The upshot, argues Kripke, is that there is no point trying to square the content of Peter's beliefs with the semantic content expressed by the embedded complements of (18)-(21), nor has it to do with the logical relations between them. Nor will it do to argue that proper names have Fregean senses or hidden Russellian descriptive content (see note 11). For in the present example, any senses or hidden content will be invariant across all type-identical name-tokens. Rather, Kripke's diagnosis is again that such *prima facie* paradoxes owe to failures of what I am calling (A)ttribution, and indirectly to (D)isquotation, which is to say instability in our belief ascription/reporting practices. And so whatever its solution, if indeed there is one, Kripke takes the "Paderewski" puzzle as showing that the apparent contradiction in Peter's beliefs cannot be blamed on failures of (S). And if (S) survives the assault then so does (M), as nowhere in the "Paderewski" puzzle is the truth of (M) presupposed.

Kripke's closet defense of (M) has not gone unchallenged, however. For instance, Sosa (1996) argues that Frege/Kripke-style puzzles can be generated without appeal to either (S) *or* (D)/(A), which presents a unique challenge to (M). By contrast, Saul (1997, 2007) claims that the same sorts of puzzles can be generated with so-called "simple" sentences (i.e., non-opacity-inducing sentences), which calls for an entirely new analysis

of the problem. Setting objections aside, whatever are the respective merits of (M) and (FR), the relevant point for purposes here is that both views are united in their assumption that whatever generates the relevant puzzles, it is not *formal* (i.e., *linguistic* or *metalinguistic*) in nature. I have briefly mentioned Frege's reasoning to this conclusion. And Kripke's "Paderewski" puzzle arguably cements the case. For again there are no *apparent* formal differences between tokens of the name 'Paderewski' to support either an (FR) or otherwise purely linguistic explanation for how a speaker can at once rationally assent to both a sentence and its negation.

In reply to the contrary, Fiengo & May (2006) argue that such anti-linguistic sentiments are grounded in misguided assumptions about the "underlying linguistic reality" of proper names. For example, Kripke assumes (and from what I gather most other philosophers of language) that the two tokens of 'Paderewski' in (22) above are actually *homophones* (i.e., 'Paderewski$_1$' and 'Paderewski$_2$'), which is to say tokens of formally distinct name-types. By contrast, while Fiengo & May (F&M henceforth) agree that the two tokens *utterances* of 'Paderewski' are tokens of distinct *expression*-types, they deny that the *name* 'Paderewski' is homophonous. Rather, F&M argue instead that the two instances of 'Paderewski' are tokens of type-distinct *expressions* of the *same name*. And since natural language grammars do not require coreference between type-distinct expressions of the same name, a speaker can rationally (albeit mistakenly) believe that the two token-*expressions* of the name 'Paderewski' in (22) do not corefer. Moreover, to learn otherwise, that is to learn the truth of (22), can be informative for a speaker who is ignorant of the fact that there is just one 'Paderewski'—i.e., just one individual who is the bearer of that name.

I will elaborate on this claim in greater detail over the next several sections, which requires substantial effort given the subtleties of F&M's proposal. But the basic moral, according to F&M, is that one cannot rule out a linguistic (or metalinguistic) solution to these puzzles in advance of an empirically motivated theory of how natural language grammars (i.e., I-languages) individuate the expressions that *they* generate. For only then can we properly understand the ways in which our *beliefs* about their identity conditions constrain our beliefs about whether two tokens of a given expression refer to the same individual (or not). Once properly understood, argue F&M, a linguistic solution not only seems plausible but vindicates Frege's earlier *Begriffsschrift* intuitions about "what's going on" in the puzzles under discussion.

6.3. *De lingua* beliefs to the rescue

To recapitulate some of what's been said to this point, it practically goes without saying that competent speakers assert declarative sentences to express beliefs, including beliefs about what other speakers believe. And while occasionally mistaken, it seems equally obvious that if their utterances are to be sincere, speakers must believe what they say. Take Max, who again by stipulation is a rational agent and competent speaker of English. When Max sincerely utters (4), repeated below, he thereby commits himself to the so-called "*de re*" belief that Cicero was a Roman statesman.[17]

(4) Cicero was a Roman statesman

For if Max did not believe what is expressed by (4) we surely would not judge his utterance sincere, even if Max happens to be mistaken about the true identity of Cicero. By (D)isquotation from above, Max's endorsement of (4) also entitles listeners to tacitly

[17] For reasons that I don't quite follow, F&M prefer the term "*non-de dicto*" to "*de re*," yet they seem to be effectively the same notions.

attribute to him belief in the proposition that it expresses. Furthermore, the principle of (A)ttribution entitles listeners to explicitly *report* Max's belief with a sentence such as (10) from above:

(10) Max believes that Cicero was a Roman statesman

Now, according to F&M, in addition to ordinary *de re* beliefs about individuals and their properties, competent speakers also have beliefs about certain properties of the *expressions* they utter in reference to those things, again to which speakers commit themselves in saying what they do. In the reverse direction, we form beliefs about the linguistic beliefs of others on the basis of what they say. Collectively, these are what F&M call *de lingua* beliefs—i.e., beliefs *about language*—to which they add "reflect fundamental aspects of our underlying linguistic competence, as well as how we employ that competence to further our communicative ends."

Among others, *de lingua* beliefs include what F&M refer to as "*de dicto*" beliefs of Assignments, where the notion of an *Assignment* is defined as follows:

> **Assignment:** An Assignment is a relation between a linguistic expression and its semantic value.

Accordingly, F&M characterize *beliefs of Assignments* thusly:

> **Beliefs of Assignments:** Beliefs of Assignment are beliefs about the semantic values of linguistic expressions.

With respect to lexical expressions, beliefs of Assignments amount to beliefs about the denotations of Words as they are used in discourse. For example, by asserting (4) above

Max not only commits himself to the *de re* belief that Cicero was a Roman statesman, he implicitly commits himself to belief in the *de dicto* Assignment expressed by (24):[18]

(24) 'Cicero' refers to Cicero

Indeed, that Max believes the proposition expressed by (24) is according to F&M constitutive of his rational use of (4), again whoever he takes the referent of 'Cicero' to be. Following F&M, I limit discussion here to proper names with the understanding that the same general principles apply to other lexical categories.

In order to properly understand the role of beliefs of Assignments and related *de lingua* beliefs, F&M's definition of an Assignment requires some unpacking, and specifically with respect to proper names. To this end, it will be helpful to first review in greater detail F&M's account of the grammatical role of proper names and the formal individuation of lexical expressions in which they occur. As we shall see, speaker-beliefs about the individuation conditions of lexical expressions influence beliefs about the individuation of Assignments in which those expressions appear. And according to F&M, the Assignments that speaker's believe are influenced by how those Assignments, and the LEs they contain, are *in fact* individuated by the speaker's internal grammar, which is to say by his/her I-language.

6.3.1. Names and their expressions relative to a discourse

As mentioned back in Chapter 2, F&M endorse a view according which names are unrepeated (non-homophonous) items listed in a speaker's lexicon and that serve as terminal nodes in syntactic trees. Following Chomsky, F&M take names, like all lexical

[18] F&M suggest that Max need not know any particular descriptive facts about Cicero to have the name 'Cicero' in his idiolect, and to use it with its conventional reference, so long as Max believes that (24) is true—that is, he must only believe (*de re*) of some individual (or other) that *he* goes by the name 'Cicero'.

items, to be individuated by certain formal properties that determine, among other things, "where a lexical item can occur in syntactic structures." Among properties that names lack, however, are *referential* properties. That is, while F&M agree that names have *bearers*, names themselves do not refer. Rather, it is only as employed in a discourse that names are used to refer unambiguously to particular objects/individuals. F&M put the claim as follows:

> To avoid confusion, we must distinguish names, which have bearers, from the occurrences of expressions containing names in discourse, which have referents. If we do, we may say that it is not really names that have reference, but linguistic expressions that occur in sentences used in discourse, which, of course, may contain names.

In structural terms, names are again taken to be terminal nodes that "occur in" or are "contained by" non-terminal *expressions of those names*. Employing standard notation, F&M use square brackets to indicate structural (i.e., syntactic) containment. For instance, the bare NPs '$[_{NP}$ Cicero]' and '$[_{NP}$ Tully]' are non-terminal lexical expressions (LEs) that contain the names 'Cicero' and 'Tully', respectively, which is to say that 'Cicero' and 'Tully' are proper constituents of the expressions '$[_{NP}$ Cicero]' and '$[_{NP}$ Tully]'.

As argued in Chapter 2, F&M agree that names combine, syntactically, with referential indices to generate indexed LEs of the form '$[_{NP}\ \alpha]_i$', where 'α' is a proper name and 'i' is a referential index for integer values > 0. Or in my more perspicuous yet somewhat more cumbersome notation, these NPs take the form '$[[_{NP}\ [_{PN}\ \alpha]][x_i]]$'. But for exegetical purposes I will stick with F&M's notation according to which '$[_{NP}$ Cicero$]_i$', for example, is an indexed LE that contains the name 'Cicero' and which in appropriate contexts can be used to refer to Cicero.

F&M also appear to adopt a Tarskian-style semantics for LEs that is similar though not identical to that detailed in my Appendix to Chapter 2. As formally they claim that an Assignment is a function *g* that maps linguistic expressions onto their semantic values. The central difference between F&M's proposal and mine is that on my view names themselves have satisfaction conditions of the form λx: *x* satisfies 'α' iff *x* is called 'α'. In turn, I would specify the contextually-determined semantic value of the indexed LE '$[_{NP}\ \alpha]_i$' as follows,

(25a) $g(i)^{\sigma} = \sigma(i),\ i > 0$

(25b) $g([_{NP}\ \alpha]_i)^{\sigma} = \lambda x$: *x* satisfies '$\alpha$' iff $x = \sigma(i)$ & $\sigma(i)$ is called by the name 'α'

where *g* is an interpretation function that assigns values to the index '*i*' relative to a Tarskian sequence, σ. Thus, according to (25b) the LE '$[_{NP}\ \alpha]_i$' is satisfied by a sequence σ just in case $\sigma(i)$ is the *i*-th object in σ and is called by the name 'α'. In this way, when the name 'α' is interpreted as occurring in an indexed LE of the form '$[_{NP}\ \alpha]_i$' it will be understood as referring uniquely and rigidly to a particular individual.

By contrast, according to F&M' names not only lack referential properties but as I understand names also lack meanings/satisfaction conditions. Rather, names for F&M are semantically empty lexical formatives that when combined with referential indices play the role of syntactically complex *variables* which do have satisfaction conditions. F&M do not spell out the details in precisely this way, but I believe (26) below is an accurate rendition of how they would specify the satisfaction conditions of what they take to be the complex variable '$[_{NP}\ \alpha]_i$':

(26) $g([_{NP}\ \alpha]_i)^{\sigma} = \lambda x$: *x* satisfies '$\alpha$' iff $x = \sigma(_i)$

Notice again that unlike my proposal the name 'α' makes no independent contribution to the satisfaction conditions of (28). So as not to misrepresent their view, however, I will for purposes of this chapter just accept F&M's assumptions along with their preferred notation.

In short, when F&M talk about Assignments as *relations* between linguistic expressions and their semantic values, I assume what they mean is something along the lines of (26). More specifically, F&M appear to think of Assignments as functions-in-*intension* (i.e., mental procedures) that generate *specifications of satisfaction conditions* on Tarskian sequences, as opposed to say *sets of sequences (or possible worlds) that satisfy expressions*. In other words, I take it that F&M embrace an I-language perspective on meaning (as opposed to an E-language perspective as this distinction was discussed in Chapter 2). However, as evidenced by the quotation above F&M also describe LEs as having referents, which requires a Church-style extension to Tarski's system.[19] Thus, I think what F&M have in mind, which matches my own intuitions, is that referring is a kind of speech act. But whatever is F&M's view on the matter, I will for convenience speak *as though* LEs have reference conditions with the understanding that *on my view* the meanings of LEs are specifiable only in terms of their satisfaction conditions. These differences notwithstanding, I agree with F&M that there are sound theoretical reasons for distinguishing names, *syntactically*, from the LEs that contain them.

6.3.2. Individuating Assignments

Consider a name such as 'Aristotle', which can be used, *inter alia*, in reference to either a particular ancient Greek philosopher (i.e., Aristotle of Stagira) or to a

[19] See Pietroski (*forthcoming*) for discussion on this point.

contemporary Greek shipping tycoon (i.e., Aristotle Onassis). Under present assumptions, these two uses of 'Aristotle' will occur as constituents, respectively, in the *non-coindexed* LEs '$[_{NP}$ Aristotle$]_i$' and '$[_{NP}$ Aristotle$]_k$'. Notice further that given standard assumptions about the relationship between coindexation and referential codependency, the grammar allows but does not require coreference between tokens of non-coindexed LEs, which is to say that these tokens can be satisfied by different Tarskian sequences. And this is to suggest that two tokens of the same name, e.g., 'Aristotle', can occur in formally distinct *Assignments* in F&M's sense. For if Assignments are relations between linguistic *expressions* and their semantic values, and Assignments are individuated by their *relata*, then any difference in either expression-type or semantic value constitutes a different Assignment. And since tokens of the LEs '$[_{NP}$ Aristotle$]_i$' and '$[_{NP}$ Aristotle$]_k$' have different semantic values, it follows that they have different Assignments.

Generally speaking, the referential indices that attach to LEs determine the formal (i.e., syntactic) identity (or difference) of their tokens. Specifically, tokens of coindexed LEs are type-identical, whereas tokens of non-coindexed LEs are type-distinct.[20] Importantly, these syntactic distinctions have semantic implications. For like the referential relationship between anaphors and their coindexed antecedents, tokens of coindexed LEs are referentially codependent; i.e., they *necessarily* corefer as dictated by the rules of grammar. On the other hand, like certain unbound (deictic) pronouns, tokens of *non*-coindexed LEs are free to corefer, or not, as determined in part by the speaker's referential intentions.

For example, consider again the names 'Cicero' and 'Tully'. While these names are customarily used in reference to the same person, nothing in the grammar *requires* their

[20] F&M (2006: 17).

coreference. Of course nothing in the grammar *prohibits* their coreference either. As such, 'Cicero' and 'Tully' will normally occur as constituents in *non*-coindexed, type-distinct LEs of the form '$[_{NP}$ Cicero$]_i$' and '$[_{NP}$ Tully$]_k$', where their coreference (or non-coreference) in a given context is governed by the speaker's referential intentions together with local naming conventions. Yet according to F&M, the names 'Cicero' and 'Tully' could in principle appear in coindexed expressions, as in '$[_{NP}$ Cicero$]_i$' and '$[_{NP}$ Tully$]_i$', and in which case their tokens would necessarily corefer. But this also implies that despite surface appearances '$[_{NP}$ Cicero$]_i$' and '$[_{NP}$ Tully$]_i$' are in fact tokens of type-*identical* LEs. Likewise, nothing precludes LEs that contain tokens of the same name-type from being non-coindexed, such as '$[_{NP}$ Aristotle$]_i$' and '$[_{NP}$ Aristotle$]_k$'. That is, while on the surface '$[_{NP}$ Aristotle$]_i$' and '$[_{NP}$ Aristotle$]_k$' *appear* to be tokens of the same expression-type, they are treated by the grammar as type-distinct because of their non-coindexation. And since '$[_{NP}$ Aristotle$]_i$' and '$[_{NP}$ Aristotle$]_k$' are non-coindexed LEs they are free to corefer, or not, as again determined by speaker-intentions and local naming conventions. However, from the perspective of successful communication it would be otiose for a speaker to use two type-distinct LEs that contain tokens of the same name-type in reference to the same individual in the same discourse (more on this below).

The relevant claim, in short, is the fact that coindexed LEs contain what might on the surface appear to be tokens of type-distinct *names* has no bearing on the formal identity/individuation conditions of the LEs in which those names occur. Rather, on F&M's account expression identity (or difference) is determined solely on the basis of whether the LEs in question are coindexed or non-coindexed. While the claim that tokens of coindexed LEs are necessarily type-identical has certain unexpected (and perhaps

unwelcome) consequences, it is not entirely baseless. For instance, in light of the discussion in Chapter 2 on phonological variation (§2.4.1.1 in particular), it seems quite possible, as F&M suggest, that some speakers might treat phonologically distinct name-pairs such as 'New York' and 'Noo Yawk', or 'London' and 'Londres', as dialectical variants of the same name-types. In turn, such speakers will both use and interpret LEs that contain these names as being coindexed and thus type-identical, as in $[_{NP}$ New York$]_i$/$[_{NP}$ Noo Yawk$]_i$, $[_{NP}$ London$]_i$ / $[_{NP}$ Londres$]_i$, and so forth.

Now, syntactic indices are not standardly used to track expression identity but rather *referential codependency* between, say, anaphors and their antecedents. F&M argue, however, that by interpreting referential indices as reflecting identity (or difference) of expression-type, the relevant semantic facts about anaphora fall out naturally. Yet theoretical elegance here comes with a cost. As using a single index to capture both sets of facts depends on some rather unorthodox commitments, including F&M's claim that nominal expressions have multiple pronominal forms. For example, the suggestion is that in a sentence such as,

(27) Oscar loves Sally because he is a good son

the LEs '$[_{NP}$ Oscar$]_1$' and '$[_{NP}$ he$]_1$' are in fact tokens of type-identical LEs despite their radically different spellings/pronunciations.[21] Yet this is a bullet that F&M are reportedly willing to bite. For if ordinary speakers are willing to tolerate other "shape transformations" such as between 'London' and 'Londres', or say between two tokens of the numeral '2' in the expression '2 + 2 = 4', then on what basis should we deny that that '$[_{NP}$ Oscar$]_i$' and '$[_{NP}$ he$]_i$' in (27) are tokens of type-identical LEs?

[21] By way of comparison, David Kaplan (1990, fn.3) observes that "Linguists seem to think that "John admires John" and "John admires Jane" have the same syntactical form and differ only in what they call "lexicalization" (though they claim that "John admires himself" differs from both syntactically)."

As we shall see below, this distinction is important to F&M's rebuttal to Kripke's *A Puzzle About Belief* and thus worth dwelling on a moment longer. Specifically, F&M describe their view as follows:

> We have also tolerated another sort of transformation, allowing that a given expression-type may have both pronominal and nonpronominal tokens; "[NP Oscar]$_1$" has its pronominal counterpart "[NP he]$_1$," so that "Oscar loves Sally because he is a good son" is of the same order as "Oscar loves Sally because Oscar is a good son."

To which they add:

> […] which form an expression takes will very much matter when figuring how tokens of an expression can be distributed in a syntactic structure. This is the import of *Binding Theory*, which can be couched as a set of constraints on the possible arrays of indices, and thus the possible arrays of expression-types, that can be manifest in phrase-markers.

The question then becomes this: If, for example, '[$_{NP}$ Oscar]$_1$', '[$_{NP}$ he]$_1$', '[$_{NP}$ himself]$_1$', and '[$_{NP}$ his]$_1$' are all phonological variants of the same expression-type, then why aren't they freely interchangeable in (28) and (29)?

(28) Oscar loves Sally because he is a good son

(29) #He loves Sally because Oscar is a good son

While F&M do not address this question, the idea seems to be that binding principles are nonetheless sensitive to otherwise tolerable "shape" differences between tokens of type-identical LEs as spelled out by the phonological system. In one of their clearest statements on the matter, F&M contend that:

> Technically speaking, there is no theory of anaphora per se; rather there is a theory that deals in sameness or difference of expressions, and their attendant interpretations. "Anaphora" is a term of art… Now, in our view, by virtue of the coindexing in (1),
>
> (1) Oscar$_1$ kissed his$_1$ mother.

> it is part of that sentence's linguistic meaning that the name-expression and the pronoun corefer, and this will be so for any utterance of that sentence. Speakers who use a sentence with coindexing as in (1) are committed to coreference simply by virtue of the form of the sentence. By using a sentence with coindexing, speakers cannot mean to express anything but that the coindexed expressions corefer. This is not at the discretion of the speakers. Speakers who use a sentence such as (1), therefore, would intend by their utterance to make a statement in which the name and the pronoun corefer. There is no other option; coreference is forced by grammar.

By way of summary, they add:

> *Our view* is the following: coreference may be grammatically determined, but not noncoreference, and this reflects the underlying structure of syntactic expressions. It is this structure that is represented by our notation, coindexing indicating that there are tokens of the same expression, noncoindexing, of different expressions. So, we distinguish (1):
>
> (1) Oscar_1 kissed his_1 mother.
>
> from (2):
>
> (2) He_1 kissed Oscar_2's mother.
>
> where the use of numerals provides a clear formal means of distinguishing occurrences of the same index from those of others, and hence tokens of the same expressions, as opposed to tokens of different ones. From this alone we can conclude that "[Oscar]" and the pronoun in (1) have the same reference. (Note that there is no particular conceptual problem with grammar determining coreference through identity; this is presumably the effect of the morpheme "self" in reflexive constructions, as in "Oscar saw himself.") No such conclusion as this follows for (2), however. Noncoindexing does not mean noncoreference; in fact it means nothing as far as reference is concerned. It only means that there are tokens of distinct syntactic types, just as coindexing means just that there are tokens of the same type.

Moreover, F&M assert that "we obtain a *better* theory of anaphora if we assume that syntactic tokens have types along the lines we have indicated." I won't quarrel over this claim, for it is presented as an empirical hypothesis.

6.3.3. Beliefs of Assignments

The upshot of the discussion thus far is that Assignments are individuated not only by their semantic values but also by the syntactic identity (or difference) of the expressions they contain. Moreover, the individuation of Assignments influences what F&M again call *de dicto* beliefs of Assignments, which to repeat are characterized as follows:

> **Beliefs of Assignments:** Beliefs of Assignment are beliefs about the semantic values of linguistic expressions.

With respect to LEs that contain proper names, beliefs of Assignments amount to beliefs about their contextually-specified denotations. By way of example, recall Max who by assumption believes that Cicero was a Roman statesman and would thereby readily assent to (4):

(4) Cicero was a Roman statesman

To reiterate an earlier claim, by asserting (4) Max not only commits himself to the *de re* belief that Cicero was a Roman statesman, he implicitly commits himself to belief in the *de dicto* Assignment expressed by (24) from earlier:

(24) 'Cicero' refers to Cicero

That Max believes the proposition expressed by (24) is again according to F&M constitutive of his rational use of (4), even if he happens to be mistaken about the true identity of Cicero.

Beliefs of Assignment are in turn governed by a close relative to (D)isquotation from above that F&M call the *Assignment Principle* (AP):

> **(A)ssignment (P)rinciple:** To be sincere, if a speaker uses a sentence containing an occurrence of the expression '$[_{NP}\ \alpha]_i$', the speaker believes an '$[_{NP}\ \alpha]_i$'-Assignment.[22, 23]

As before, '$[_{NP}\ \alpha]_i$' is an indexed NP (i.e., nominal LE) that contains the proper name 'α'. And like both (D)isquotation and (A)ttribution from above, that competent speakers are at least tacitly aware of (AP), and generally adhere to it, is according to F&M an aspect of their basic linguistic competence. More specifically, F&M claim that (AP) follows from two things: "the desire on the part of the speakers to speak truthfully, and the fact that it is a condition on the truth of sentences that the expressions have the appropriate values."

Now, not only do competent speakers have beliefs of Assignments, they have beliefs about the Assignments that other speakers believe, again as justified by what those speakers say. With respect to proper names, F&M state that "not only do we know what names we give to people; we know what names other people give to people." And this is to say that competent speakers have what I will call *beliefs of Attributed Assignments*:[24]

> **Beliefs of Attributed Assignments** are beliefs about the Assignments that other speakers believe and which can be justifiably attributed to them as governed by (AP), which is to say as evidenced by their verbal behavior.

When speakers verbally express their beliefs of Attributed Assignments they are in effect reporting their beliefs about another speaker's beliefs of Assignment, again as justified by (AP). In the present example, the report will take the form of a sentence like (30):

(30) Max believes that 'Cicero' refers to Cicero

---

[22] They add that (AP) follows from two things: "the desire on the part of the speakers to speak truthfully, and the fact that it is a condition on the truth of sentences that the expressions have the appropriate values."
[23] Notice that I am altering F&M's notation slightly to remain consistent with my own, yet without loss of detail or accuracy.
[24] F&M do not formulate what I am calling Beliefs of Attributed Assignments in so many words, but this seems to be essentially what they have in mind.

Notice, however, that (AP) does *not* license the Attribution expressed by (31) below, or at least not as based solely on Max's explicit endorsement of a sentence such as (4):

(31) Max believes that 'Tully' refers to Tully

Why not? In short, this is because 'Cicero' and 'Tully' are tokens of formally distinct (or what F&M called "non-cospelled") name-types, which *suggests* that the LEs in which these names occur are tokens of formally distinct expression-types. Max therefore believes *formally* distinct Assignments that happen to have the same semantic values. Max's utterance of (4), however, commits him only to belief in the 'Cicero'-Assignment in (24).

The Attribution of Max's 'Tully'-Assignment expressed by (32) below is however justified by his endorsement of (6):

(6) Tully was a Roman poet

(32) 'Tully' refers to Tully

For if Max believes (6) then by (AP) he is thereby committed to the belief of Assignment expressed by (32) just as it is attributed to him in (31). The upshot, according to F&M, is that "Assignments are distinguishable strictly on linguistic grounds." (*ibid*: 16). That is, while the Assignments attributed to Max in (30) and (31) have the same *semantic values* (i.e., referential content) they are nevertheless *formally* (i.e., *syntactically*) distinct Assignments because they contain tokens of formally distinct name-types.

The formal individuation of Assignments, and correlatively their Attributions, is not always clear cut, however. For suppose that Max's cousin Minnie sincerely asserts both (33) and (34).

(33) Aristotle was an ancient Greek philosopher

(34) Aristotle was a twentieth century Greek shipping tycoon

where it is clear from the context that Minnie is using the name ‘Aristotle’ in (33) and (34), respectively, in reference to two different individuals. In other words, Minnie believes two referentially distinct Assignments each involving the name ‘Aristotle’ yet which have different semantic values (*cp*., Pierre’s beliefs about the two ‘Paderewski’s from above; more on this below). Under present assumptions Minnie’s interlocutor, call him Sam, is entitled to the Attribution reported with (35):

(35) Minnie believes that ‘Aristotle’ refers to Aristotle

The question is of course *which* of Minnie’s Assignments is being ascribed/reported with Sam’s utterance of (35)—i.e., is it the Assignment whose semantic value is the philosopher or the one whose value is the tycoon? The answer here seemingly depends on who *Sam* is using the name ‘Aristotle’ to refer to. This in turn suggests that either the name ‘Aristotle’ is referentially ambiguous (i.e., homophonous) or that its respective tokens out of Sam’s mouth are not tokens of the same expression-type.

As indicated earlier, since on F&M’s view names do not have referents, the name ‘Aristotle’ cannot be referentially ambiguous. Rather they claim that contrary to appearances, in her respective utterances of (33) and (34) Minnie is in fact using formally distinct expressions of the same name—i.e., LEs that have different semantic values yet which happened to be “cospelled” (have the same phonological form) And this is again to suggest that the underlying syntactic forms of LEs are more fine-grained than their surface forms suggest. What’s more, competent speakers have *beliefs* about how LEs are formally (i.e., syntactically) individuated, or what F&M call *beliefs of Identity*:

> **Beliefs of Identity** are beliefs about the syntactic identity (or difference) of linguistic expressions.

As suggested earlier, beliefs about the individuation conditions of LEs constrain beliefs about whether or not their tokens are covalued (or in the case of names whether or not they "corefer"). Specifically, speakers tacitly know that tokens of type-identical LEs *must* be covalued as enforced by the grammar. Thus, speakers who believe of two LEs that they are coindexed will also naturally believe that they are covalued. On the other hand, speakers who believe of two LEs that they are non-coindexed will believe that they may or may not be covalued, again as determined by the speaker's referential intentions in conjunction with customary naming conventions.

In short, by assumption Minnie believes that 'Aristotle' has two distinct bearers; one a philosopher and the other a shipping tycoon. And this is to say that Minnie believes two non-coindexed, non-coreferential (or non-covalued) Assignments each containing an LE that itself contains the name 'Aristotle'. If expressions are individuated in part by their semantic values (i.e., referents), and the two 'Aristotle'-expressions in Minnie's Assignments do not corefer, then Minnie's 'Aristotle'-expressions cannot be tokens of the *same* expression-type, which is precisely what F&M contend:

> While there is only one name "Aristotle"—one phonetic type "Aristotle," one "spelling" or "shape" "A^ r^ i^ s^ t^ o^ t^ l^ e"—there are many distinct syntactic expression-types that contain the name "Aristotle."

Likewise, if Minnie *believes* two different 'Aristotle'-Assignments, she will thereby believe that those Assignments contain type-distinct 'Aristotle'-expressions (LEs) that refer to numerically distinct individuals. In general, F&M claim that:

> … a speaker will believe that there are as many expression-types containing a given name as she believes there are people bearing that name… In a given discourse, a speaker's use of expressions of a name will directly reflect the speaker's beliefs about how many values

that name has; in referring to each, a speaker will use different expressions, for only this will comport with his beliefs (otherwise his expressions would corefer).

The question of whether or not type-distinct expressions corefer is again a matter of linguistic convention, not grammar. That is, beliefs about the coreference of type-distinct expressions depend on beliefs about what other speakers use these expressions to refer to. And this is intuitively at least part of the story for why true identity statements of the form 'α is β' can be informative for competent speakers who are otherwise ignorant of the relevant naming convention. For given that the expressions '$[\alpha]_i$' and '$[\beta]_k$' are tokens of formally distinct expression-types, a speaker might mistakenly (though quite naturally) believe that they do not corefer. Indeed, while F&M do not say this, recall from Chapter 4 that in the absence of evidence to the contrary the *Mutual Exclusivity* bias naturally predisposes speakers to assume that type-distinct expressions are not covalued. And so to learn otherwise would seemingly constitute a substantive advance in the speaker's knowledge about (i) the referential properties of certain linguistic expressions, and thereby (ii) the identity and number of individuals referred to by these expressions. In other words, the question of informativeness seems to turn, in some way or other, on the fact that '$[\alpha]_i$' and '$[\beta]_k$' are tokens of formally distinct (i.e., non-coindexed) expression-types.

I will return to the topic of informativeness below and again in Chapter 7. But back to our previous example, the fact that Minnie uses the name 'Aristotle' in reference to two individuals implies that she believes two formally distinct 'Aristotle'-Assignments that contain two non-coindexed 'Aristotle'-expressions as their respective constituents. For instance, let us just stipulate, as do F&M, that the LE '$[\text{Aristotle}]_1$' refers to the philosopher and that '$[\text{Aristotle}]_2$' refers to the shipping tycoon (to avoid clutter, the

category label 'NP' shall hereafter be left implicit).[25] Thus, the (partial) logical forms of their respective Assignments in Minnie's idiolect are specifiable as follows:

(35) '$[\text{Aristotle}]_1$' refers to $[\text{Aristotle}]_1$ (= the philosopher)

(36) '$[\text{Aristotle}]_2$' refers to $[\text{Aristotle}]_2$ (= the shipping tycoon)

Notice that the second occurrence of '[Aristotle]' in both (35) and (36) is being *used* rather than mentioned. In other words, like traditional meaning axioms the second occurrence represents the semantic contribution of its first occurrence to the proposition expressed by these Assignments.

So qualified, we can specify more precisely which of Minnie's 'Aristotle'-Assignments are being reported with Sam's utterance of (34):

(34) Minnie believes that 'Aristotle' refers to Aristotle

For under the assumption that Sam's grammar can generate both 'Aristotle'-expressions, he can use (34) to report *either* Assignment in (35) or (36), depending on the logical form of his utterance, which by hypothesis will be one of the following:[26]

(37') Minnie believes that '$[\text{Aristotle}]_1$' refers to $[\text{Aristotle}]_1$

(37") Minnie believes that '$[\text{Aristotle}]_2$' refers to $[\text{Aristotle}]_2$

Precisely which of these forms his utterance takes presumably depends on what it is that Sam intends to express. And while there are open questions as to how speaker-intentions determine the logical forms of utterances, that they do is generally taken for granted.

---

[25] For their part, F&M use double quotes around names as they appear in expressions, as well as different position for indices; for example [$_1$ "Aristotle"]. However, I will continue to use single quotes to maintain consistency with notation employed in earlier chapters. Indeed, the notation employed by F&M is itself inconsistent, apparently due to the fact that *DLB* is a collection of independently written essays.

[26] F&M say very little about how a speaker chooses between logical forms, and how listeners recover these forms in a given context of utterance. Yet we seem to manage such tasks with other ambiguous expressions on a daily basis. And so for purposes here I am willing to indulge the assumption that there is a mechanism that explains how we sort such things out.

Let me mention in passing that while it may again strike one as obvious that '[Cicero]' and '[Tully]' are type-distinct LEs in virtue of containing tokens of non-cospelled name-types, theoretical consistency dictates that they too carry syntactic indices. So for completeness let us represent the logical forms of their Assignments as follows:

(24') '$[\text{Cicero}]_1$' refers to $[\text{Cicero}]_1$ (= Marcus Tullius Cicero)

(32') '$[\text{Tully}]_2$' refers to $[\text{Tully}]_2$ (= Marcus Tullius Cicero)

Here again the index-values that we as theorists assign to these expressions is arbitrary. What matters for purposes of interpretation in this case is that they differ. Also keep in mind that non-coindexation neither entails nor precludes non-coreference—it allows for either, with one exception to be discussed presently. Yet one assumes that there is a fact of the matter, and that in the present case I am assuming that the expressions '$[\text{Cicero}]_1$' and '$[\text{Tully}]_2$' do in fact corefer.

6.3.4. Coindexation, coreference, and Singularity

This last caveat is actually quite important, for while non-cospelled (i.e., formally distinct) names are *indicative* of a difference in *expression*-type, on F&M's view, as mentioned earlier, a difference in name-type does not *guarantee* a difference in expression-type. For consider again the following coindexed expression pairs: '[New York]$_1$'/'[Noo Yawk]$_1$' and '[London]$_1$'/ '[Londres]$_1$', which are plausibly phonological variants of the same expression. But even if speakers do not treat non-cospelled expression-tokens as type-identical, they may nevertheless believe that they are what F&M call *Translations* of each other. Correlatively, such speakers have what F&M refer to as *beliefs of Translation*:

> **Beliefs of Translation** are beliefs that non-coindexed (type-distinct), non-cospelled expressions corefer.

For example, I take it that most philosophers and historians believe that the type-distinct expressions '[Cicero]$_1$' and '[Tully]$_2$' are Translations of each other, but that, say, '[Cicero]$_1$' and '[Cataline]$_2$' are not. Beliefs of Translation actually play a pivotal role in F&M's solution to Frege's puzzle of informativeness, which I will get to momentarily.

By contrast, in the case of type-distinct yet cospelled expressions such as '[Aristotle]$_1$' and '[Aristotle]$_2$', no speaker, as F&M (*ibid*: 147) put it, would ever assume that "one person can have one particular name more than once," which rules out coreference of the *LEs* '[Aristotle]$_1$' and '[Aristotle]$_2$' in typical circumstances. That is, this situation is the exception to the general rule that type-distinct expressions *optionally* corefer. The exception here is governed by a principle that F&M call *Singularity*, which I abbreviate as **(SI)** to avoid confusion with the principle of (S)ubstitutivity from above:

> **(SI)ngularity:** Speakers believe that cospelled expressions corefer if coindexed, and if non-coindexed that they do not corefer.

Unfortunately, F&M are unclear as to whether (SI) is governed by the grammar or whether it is extralinguistic/pragmatic in nature (I take the latter to be the intended view). Whatever is the case, as we shall see momentarily (SI) is central to F&M's solution to "Padereswki"-style puzzles. For (SI) is the principle that serves to bridge the epistemic gap between (i) a speaker's *de re* beliefs about how many individuals are the bearers of a given name, and (ii) his or her *de dicto* beliefs about whether or not two names corefer. Before proceeding to the puzzles, let me take a moment to rehearse one last assumption that is equally central to F&M's account, and arguably the most controversial.

6.3.5. Expressing beliefs of Assignments

Recall that beliefs of Assignments can be verbally attributed (and reported) as licensed by the *Assignment Principle* (AP), repeated here for convenience:

> **(A)ssignment (P)rinciple:** To be sincere, if a speaker uses a sentence containing an occurrence of the expression '$[_{NP}\ \alpha]_i$', the speaker believes an '$[_{NP}\ \alpha]_i$' Assignment.

For instance, Max's belief of Assignment expressed by (1'), repeated below, can be reported with (29) whose (partial) logical form is given in (29'):

(24') '$[\text{Cicero}]_1$' refers to $[\text{Cicero}]_1$

(30) Max believes that 'Cicero' refers to Cicero

(30') Max believes that '$[\text{Cicero}]_1$' refers to $[\text{Cicero}]_1$

Now, vital to the account developed in *DLB* is that not only can listeners overtly report Max's belief of Assignment with (30'), the Assignment that it attributes to him can in the right circumstances be *optionally expressed* as a *covert* part of the proposition expressed by an utterance of (10):

(10) Max believes that Cicero was a Roman statesman

Specifically, the hypothesis is that the logical form of (10), as uttered in discourse, is systematically ambiguous as between (10') and (10''):

(10') Max believes that [$[\text{Cicero}]_1$ was a Roman statesman]

(10'') Max believes that [$[\text{Cicero}]_1$ was a Roman statesman] & ['$[\text{Cicero}]_1$' refers to $[\text{Cicero}]_1$]]

That is, (10) can be used to ascribe to Max *either* an ordinary *de re* belief about Cicero as analyzed in (10') or this same belief *optionally conjoined* with a *de dicto* belief of

Assignment as in (10").[27] Thus, F&M suggest that sentences such as (10) have both *de dicto* and what they call "non-*de dicto*" logical forms. Specifically, (10') is the non-*de dicto* logical form of (10), and (10") is its *de dicto* logical form. In general, F&M claim that the non-*de dicto* logical form of a sentence *expresses* its *objectual content*, whereas its *de dicto* logical form *conveys* in addition its *informational* or *linguistic content*. As we shall see below, *de dicto* logical forms are on F&M's view what explain the informativeness of identity statements containing type-distinct coreferential names of the form '$[a]_i = [b]_j$' (or in natural language, '$[\alpha]_i$ is $[\beta]_j$').

Importantly, F&M say of *de dicto* logical forms such as (10") that "*one* belief is attributed, that one belief being expressed as the conjunction of two sentences."[28] The "one belief" remark is crucial because it implies that the believer (in this case Max) has *integrated* (i.e., brought together in his mind) both conjuncts of the belief attributed to him in (10") under a single thought. And this is to suggest that the attribution in (10") above is *not* that expressed by (10"') below:

(10"') Max believes that [[Cicero]$_1$ was a Roman statesman] & Max believes that ['[Cicero]$_1$' refers to [Cicero]$_1$]

For according to F&M (10"') can be true if Max believes its two conjuncts independently, "having never considered the two beliefs together," a possibility that is reportedly excluded by (10"). Whatever one makes of this last claim, the upshot is that:

> … when certain conditions are met, the content of beliefs about the reference of expressions can be taken to be part of what we say by our utterances, part of the propositional content; to put it a little differently, they can be part of the *interpretation* of

[27] F&M (2006: 64) state that "It is natural to call attributions containing Assignments *de dicto*; their logical forms explicitly contain the expressions with respect to which the beliefs are held. More specifically, (10") is what F&M regard a "disquotational" *de dicto* logical form for reasons I take to be apparent.
[28] My emphasis. F&M suggest (*ibid*), in addition, that for the attribution to be *correct*, it must be the case that the agent would agree that it is Cicero is being referred to by the 'Cicero'-expression.

> our utterances, by ourselves and by others. By "part of" here we mean *formal* part—that is, as constituents of the logical form that expresses this interpretation, that represents what we say.

If correct, however, one still wants to know *why* a speaker might attribute a belief of Assignment, covertly or otherwise. I will also say more about this below, but one reason derives from the fact that beliefs of Assignment are idiosyncratic—i.e., they vary from speaker to speaker. An Assignment might therefore be attributed in order to draw an audience's attention to the fact that a particular expression is being used as the *subject* of the report uses it, and thus not (necessarily) as the reporter uses it.[29] In other words, beliefs of Attributed Assignments are expressed when a reporter wishes to say of the agent of the report that he/she believes something with respect to a particular Assignment. In F&M's words:

> The speakers, for example, may wish to speak of the terms under which an agent holds a belief. In that case, the speakers may specify the particular Assignment that they believe the agent would assent to. The speakers may, that is to say, attribute Assignments.

However, doing so covertly is only likely to be effective when both the reporter and his/her audience know that:

> … there is some *other* attribution, containing some other equivalent Assignment, that the agent does not believe. If that constitutes the reason for the inclusion, then it stands as an *implicature* of an attribution containing an Assignment that there is some other incorrect attribution, an attribution containing an Assignment that the agent does not believe.

[29] In this case, the attributed Assignment is being expressed with the reporter's intention of notifying the subject of a divergence in their respective beliefs about the identity of particular linguistic expression. In turn, the *linguistic* information expressed by the attribution, if recognized, may be sufficient to cause the subject's to bring his *objectual* beliefs about how many individuals are the bearers of a particular name into conformity with those of the reporter.

A second reason is to draw the subject's attention to the fact that two type-distinct (i.e., non-coindexed) expressions corefer or, conversely, that they do not corefer.[30] As we shall see in a moment, this second condition is essentially what grounds F&M's proposed solution to both Frege's puzzle about the informativeness of identity statements and Kripke's "Paderewski" puzzle about failures of substitution in propositional attitude reports.[31]

In sum, the key to understanding both puzzles, claim F&M, is to recognize that they turn on two kinds of beliefs held by speakers and hearers. The first are *de re* (objectual) beliefs about how many individuals, *x*, are the bearers of a particular name, 'α'. The second are the *linguistic* beliefs just surveyed, which is to say *de dicto* beliefs of *Identity*, *Assignment*, *Attributed Assignment*, *Translation*, and *Singularity*. Disagreement (or confusion) between speakers and listeners over such beliefs are what *generate* the relevant puzzles, whereas the *de dicto* logical forms of sentences used to report these beliefs are reportedly what *dissolve* them, as I attempt to demonstrate next. I begin with Frege's puzzle about the informativeness of identity statements followed by F&M's related analysis of Kripke's "Paderewski" puzzle about belief sentences. Thereafter I recapitulate what appear to be two devastating objections that illustrate why, despite otherwise sound intuitions, *DLB* fails to achieve its stated goals.

[30] The reporter is again entitled to such attributions only under the conditions specified by (AP). As will also become clearer below, according to F&M it is *apparent* exceptions to (AP) that generate the puzzles about belief sentences.

[31] As an advertisement, F&M suggest that when properly framed the latter is just "a new puzzle about identity statements." As mentioned in the beginning, I happen to agree.

6.4. DLB's solution to the puzzles

6.4.1. Frege's puzzle of informativeness

Frege's puzzle of informativeness again amounts to this: Why should the substitution of type-distinct, non-cospelled, coreferring proper names in an identity statement alter the informativeness or "cognitive significance" of the thought that it expresses? I again take it that all parties agree that identity statements of the form 'a = b' (in formal languages) or 'α is β' (in natural language) are (or can be) *informative* for subjects who understand them, where the relevant notion of informativeness I take to be governed, more or less, by the following principle:

> **(INF)ormativeness:** A true identity statement of the form 'α is β', where 'α' and 'β' are lexical expressions that contain coreferential uses of proper names, is *informative* just in case its utterance expresses or otherwise pragmatically conveys information that is sufficient to rationally compel a linguistically competent yet otherwise misinformed subject who understands its meaning, and believes it true, to revise his/her mistaken beliefs about *how many* objects are the bearers of those names. Parallel remarks apply to statements of the form 'α is *not* β'.

The trick is again to specify (i) *what*, exactly, is the nature of the information conveyed by 'α is β' (or 'a = b') that renders it informative, and (ii) *how* does this information (causally) effect relevant changes in the subject's system of beliefs.

In the first instance, as you might have guessed, the relevant information is according to F&M *linguistic* in nature (or again what I would call *metalinguistic*). More specifically, it has to do with what they call the "informational content" conveyed by the *de dicto* logical forms of sentence-utterances. Broadly speaking, F&M (2006: 113) say:

> Let us identify the *informational content* of a sentence with the entailments of its logical form.[32] A sentence, by virtue of having the logical form that it does, *has* informational content; a statement, we then say, *conveys* the informational content the sentence used to make that statement has.

I will clarify these remarks in a moment, but in the second instance the informational content carried by such statements is claimed to be sufficient to bring about relevant changes in a subject's beliefs by first rationally compelling him/her to reconsider his/her standing *de dicto* beliefs about whether or not the two names in question corefer. If accepted, this information will provide the subject with a crucial premise that logically connects his/her *de dicto* beliefs of Assignments to his/her *de re* or *objectual* beliefs about how many individuals are referred to in those Assignments. If that connection is properly grasped, the resulting inference will in turn provide grounds for bringing both kinds of beliefs into alignment with those of the subject's informant.

The reasoning just sketched is an example of what F&M mean by "entailments of logical form." Recall (10) from above, whose logical form is *ex hypothesi* ambiguous as between (10') and (10"):

(10') Max believes that [[Cicero]$_1$ was a Roman statesman]

(10") Max believes that [[Cicero]$_1$ was a Roman statesman] & ['[Cicero]$_1$' refers to [Cicero]$_1$]

(10') again represents the "non-*de dicto* logical form" of (10), which is to say that it specifies only the *objectual* content of (10); i.e., *sans* the *de dicto* Assignment in (10"). However, the non-*de dicto* logical form of (10) *entails* that its *de dicto* logical form in (10") will include an Assignment involving the expression '[Cicero]$_1$'. Naturally, the first

[32] F&M (*ibid*: 113) add: "Let us use the term *logical form* to designate that part of the syntactic description of a sentence that determines (along with other nonlinguistic factors of context of utterance and indexicality that we set aside), the truth conditions of statements made by using that sentence."

conjunct of (10") expresses the same *objectual* content as (10'). But it is the *linguistic* (i.e., *de dicto*) content of the Assignment in (10") that F&M identify with its *informational* content.

Identity statements are reportedly no different in this respect.[33] For instance, (3) from earlier, repeated below, is ambiguous as between the non-*de dicto* logical form of (3') and its *de dicto* logical form in (3") which in this case contains three conjuncts:

(3) Cicero is Tully

(3') $[\text{Cicero}]_1$ is $[\text{Tully}]_2$

(3") $[[\text{Cicero}]_1 \text{ is } [\text{Tully}]_2] \ \&\ (\text{`}[\text{Cicero}]_1\text{'} \text{ refers to } [\text{Cicero}]_1) \ \&\ (\text{`}[\text{Tully}]_2\text{'} \text{ refers to } [\text{Tully}]_2)]$

Here again (3') and the first conjunct of (3") express the same objectual content, or as Frege would say they express the same "objectual identity." Moreover, the *objectual* content of (3) is identical to that expressed by (1") and (2") below, which are of course both trivially true.

(1") $[\text{Cicero}]_1$ is $[\text{Cicero}]_1$

(2") $[\text{Tully}]_2$ is $[\text{Tully}]_2$

Indeed even the informational contents of (1") and (2")—i.e., their *de dicto* logical forms—in isolation are uninformative, which are those in (1'") and (2'"):

(1'") $[\text{Cicero}]_1$ is $[\text{Cicero}]_1$ & $\text{`}[\text{Cicero}]_1\text{'}$ refers to $[\text{Cicero}]_1$

(2'") $[\text{Tully}]_2$ is $[\text{Tully}]_2$ & $\text{`}[\text{Tully}]_2\text{'}$ refers to $[\text{Tully}]_2$

The *de dicto* Assignment in (1'"), for example, conveys nothing more than the fact that the two occurrences of $\text{`}[\text{Cicero}]_1\text{'}$ in the first conjunct refer to the same thing, which

[33] More generally, F&M claim that *all* sentences are ambiguous between having *de dicto* and non–*de dicto* logical forms.

carries the trivial non-*de dicto* entailment that the individual so-referred is self-identical. Parallel reasoning applies to (2'").

By contrast, because (3) contains type-distinct names its *de dicto* logical form in (3") will contain *two* Assignments, one for each expression of those names. The difference is relevant because (3") "includes," as F&M put it, the "information" that '$[\text{Cicero}]_1$' and '$[\text{Tully}]_2$' corefer. For given the collective truth of the three conjuncts in (3"), F&M contend that their coreference "follows trivially." Specifically, they reason that (3"):

> … entails that the reference of the expression "Cicero" is the same as the reference of the expression "Tully," for if "Cicero" refers to Cicero, and "Tully" refers to Tully, and Cicero and Tully are one and the same, then "Cicero" and "Tully" corefer. *De dicto* logical forms convey information by virtue of what they *show* or *display* that their paired non–*de dicto* counterparts do not.

The idea seems to be that if an utterance of (3) is understood as expressing (3"), the result is a single thought that integrates all three conjuncts. The information conveyed by its *de dicto* logical form is therefore what renders an utterance of (3) *informative*, as recognition of the thought that it expresses will, *ceteris paribus*, be sufficient to cause Max to adjust his *de re* beliefs about how many individuals are the bearers of the names 'Cicero' and 'Tully'. This is all just to reinforce Fregean intuitions, couched in linguistic terms, that an utterance of (3) is (or at least can be) informative in a way that (1) and (2) are not.

I will explain more precisely how this is supposed to work in a moment, but notice that the same clearly cannot be said of (3'). For the informational content conveyed by (3') is nothing over and above what is apparent from its surface form, which is that the expressions '$[\text{Cicero}]_1$' and '$[\text{Tully}]_2$' are non-coindexed (i.e., type-distinct) and, hence,

that their tokens *optionally* corefer, and which with respect to (INF) from above is *uninformative*. To this point F&M add:

> A speaker who makes an utterance of "Cicero is Tully" does so backed by the beliefs that (i) Cicero is Tully, (ii) "Cicero" has the value Cicero, and (iii) "Tully" has the value Tully. *Only if she utters a sentence with a de dicto logical form do the latter beliefs become part of what is said by the speaker*; by virtue of the explicit occurrences of Assignments, she will convey the information that two distinct expressions corefer. [my emphasis].

The putative entailment conveyed by (3") is perhaps more easily seen in the following logically equivalent regimentation; I will call this '(P3)' for reasons that will become clear shortly:

(P3) $\exists x \exists y$ [(x is referred to by '$[\text{Cicero}]_1$') & (y is referred to by '$[\text{Tully}]_2$') & (x=y) $\supset$ '$[\text{Cicero}]_1$' and '$[\text{Tully}]_2$' corefer]

So, we now know *what* information putatively accounts for the informativeness/cognitive significance of identity statements. However, it remains to be said precisely *how* this information effects the relevant changes in a listener's beliefs. Let's take the case of Max first, which we might think of as the *easy* case involving the belief that non-cospelled (type-distinct) names have numerically distinct bearers. The more difficult case to explain, which I turn to below, involves one *name*, such as 'Aristotle' or 'Paderewski', which is believed to have multiple bearers.

By hypothesis, Max, like other competent speakers, knows/believes that '$[\text{Cicero}]_1$' and '$[\text{Tully}]_2$' are non-coindexed and thus may optionally corefer. And in the absence of evidence to the contrary, it should again come as no surprise that Max might believe that '$[\text{Cicero}]_1$' and '$[\text{Tully}]_2$' in fact do not corefer. Thus, we can say that Max mistakenly but rationally believes (P2):

(P2) $\exists x \exists y$ [(‘[Cicero]$_1$’ refers to x) & (‘[Tully]$_2$’ refers to y) & ($x \neq y$)]

The question is how can an utterance of (3), which expresses only trivial objectual content, cause Max’s *de re* beliefs to become coincident with those of its speaker? This is where the informativeness of (3") comes into play, for it putatively conveys to Max the crucial “bridge premise” specified by (P3) above, which is that ‘[Cicero]$_1$’ and ‘[Tully]$_2$’ corefer. In other words, (3") reportedly conveys the information that these two expressions are *Translations* of each other. Max’s acceptance of (P3) will thereby provide him with grounds to adjust his *de re* beliefs about how many individuals are the bearers of the names ‘Cicero’ and ‘Tully’.

Even supposing that Max grasps the proposition expressed by (P3), one still might wonder why he might be compelled to accept it. According to F&M (*ibid*: 117), this is because “ordinarily we speak assertively (and sincerely) just because we want to speak informatively.” Max will therefore assume that his interlocutor is using (3) to tell him something that he doesn’t know. F&M (*ibid*: 23) add that “if a hearer properly understands what a speaker says, he or she will come to represent the sentences the speaker utters as the speaker does.” And since Max knows that the only contingent fact about the logical form of (3") is whether or not ‘[Cicero]$_1$’ and ‘[Tully]$_2$’ corefer, he will be compelled to consider the possibility that his informant has included the content expressed by (3")/(P3) specifically to challenge his disbelief that the Assignments that it embeds are covalued.[34] Should Max defer to the challenge, which is to say should he (i) manage to extract its intended content, and (ii) find that content persuasive, he will thereby feel rationally obliged to align his (*de re*) beliefs with those of his informant about how many individuals are the bearers of the names ‘Cicero’ and ‘Tully’. And this,

[34] F&M don’t put it quite this way, but I take it this is what they have in mind.

argue F&M (*ibid*: 104), is "how identity statements can be used to tell someone that names are coreferential, so that they come to be *informed*," which in turn demonstrates one way in which "linguistic information can be semantically significant" (*ibid*: 142).

If some of this strikes you as a bit fishy, you are in good company. But hold onto that thought. For under that assumption that what has been said here is coherent, it becomes a relatively simple task to show how F&M's proposal extends to Kripke's "Paderewski puzzle," which is just a special case of (INF)—i.e., the informativeness of identity statements involving coreferential expressions.

6.4.2. The Paderewski puzzle

First recall Kripke's Peter, who mistakenly believes that there are two individuals with the same name; one a politically talentless musician and the other a musically talentless politician. In consequence of his confusion, Peter assents to both (15) and (16):

(15) Paderewski lacks musical talent

(16) Paderewski has musical talent

Next recall F&M's principle of *Singularity*, repeated below:

(SI)ngularity: Speakers believe that cospelled expressions corefer if coindexed, and if non-coindexed that they do not corefer.

Now, by (SI) coupled with Peter's belief that there are two individuals named 'Paderewski', he will thereby believe that there are *two non-coindexed, cospelled expressions of that name*. That is, by assumption Peter believes the following Assignments:

(38) '$[\text{Paderewski}]_1$' refers to $[\text{Paderewski}]_1$

(39) '$[\text{Paderewski}]_2$' refers to $[\text{Paderewski}]_2$

Thus, (15) and (16) *out of Peter's mouth* will have the following (non-*de dicto*) logical forms:

(15') [Paderewski]$_1$ lacks musical talent

(16') [Paderewski]$_2$ has musical talent

Given (SI), Peter will thereby reject the identity expressed by (22), whose logical form is given in (22'):

(22) Paderewski is Paderewski

(22') [Paderewski]$_1$ is [Paderewski]$_2$

For according to Peter's idiolect, (22) has the form 'α is β'. And again given his beliefs about Singularity, Peter will believe that his two 'Paderewski'-expressions *do not* corefer. And so in this way Peter can rationally assent to (16) while denying (15).

By the same token a duly informed listener, call her Penny, can infer from Peter's verbal behavior his beliefs of Assignments in (38) and (39)—i.e., his belief that there are two cospelled, non-coindexed expressions, '[Paderewski]$_1$' and '[Paderewski]$_2$', that do not corefer. In turn, by (AP) from above Penny is entitled to the following attributions without thinking Peter irrational:

(20') Peter believes that [Paderewski]$_1$ has musical talent

(21') Peter does not believe that [Paderewski]$_2$ has musical talent

For only under this condition would Penny not be attributing to Peter the contradictory belief of the form *p* & *not-p*. As F&M (*ibid*: 22) describe the situation:

> If the occurrences of "Paderewski" are not occurrences of the same expression-type, then it does not follow, as a matter of logical form, that inconsistent beliefs have been attributed to the agent, for it is left open grammatically whether such occurrences corefer or not.

So, the question now becomes: What could be said to relieve Peter of his confusion about the true identity of Ignacy Jan Padereswki? In other words, what premise might be sufficient to alter Peter's mistaken beliefs about the number of bearers of the name 'Paderewski'? Since Penny believes that there is just *one individual* named 'Paderewski', it follows from (SI) that *she* believes there is just *one expression* containing that name, say '[Paderewski]$_1$'. As such, Penny *cannot* sincerely assert (22) with the non-*de dicto* logical form of (22'), for this is not consonant with her beliefs. Penny can however assert (22) with its *de dicto* logical form as specified in (22"):

(22") [(Paderewski]$_1$ is [Paderewski]$_2$) & ('[Paderewski]$_1$' refers to [Paderewski]$_1$) & ('[Paderewski]$_2$' refers to [Paderewski]$_2$)]

For in this case Penny is free to use the name 'Paderewski' *as Peter uses it*. What's more, as F&M suggest (*ibid*: 77), she is saying to *him*, in effect, that "*your* "Paderewski"-expressions corefer."

I take it that F&M's qualification "in effect" above is purposeful. For as we have seen, (SI) prohibits beliefs of coreference between cospelled, non-coindexed (type-distinct) names. And since Penny and Peter presumably both adhere to (SI), they both know/believe that the two 'Paderewski'-expressions in (22") cannot corefer. And so, more accurately we should say, as F&M do elsewhere (*ibid*: 77), that what Penny is trying to tell Peter is that "two of *his* Assignments are equivalent." And this, I take it, will *convey* to Peter the message that there is only *one* 'Paderewski'-*expression* and that *it* refers to *the* man Ignacy Jan Padereswki. This epiphany will, as before, trigger a chain of reasoning that rationally compels Peter to bring his *de dicto* beliefs of Assignments into conformity with those held by Penny.

In a more regimented formulation, Peter's chain of reasoning is (roughly) an instance of the following inference scheme, where (P3)—the crucial premise—is isomorphic to the *de dicto* logical form of Penny's utterance of (22"):

(P1)	$\exists x \exists y$, $\mathbf{x \neq y,}$ [(x ***is named*** ‘a’) & (y ***is named*** ‘α’)]	Assumption
(P2)	$\exists x \exists y$, $\mathbf{x \neq y,}$ [(‘$[\alpha]_1$’ ***refers to*** x) & (‘$[\alpha]_2$’ ***refers to*** y)]	(P1), (SI)
(P3)	$\exists x \exists y$, $\mathbf{x = y,}$ [(‘$[\alpha]_1$’ ***refers to*** x) & (‘$[\alpha]_2$’ ***refers to*** y)]	(P2)
(P4)	$\exists x \exists y$, $\mathbf{x = y,}$ [(x ***is named*** ‘α’) & (y ***is named*** ‘α’)]	(P3), (SI)
(P5)	$\exists x$ (x ***is named*** ‘α’)	(P4), Simp
(P6)	$\exists x$ [(x ***is referred to*** by ‘$[\alpha]_1$’) & (x ***is referred to*** by ‘$[\alpha]_2$’)	(P5), (SI)

With assumptions discharged, F&M take themselves to have provided a *linguistic* solution to Kripke's "Paderewski puzzle" though framed specifically as a puzzle about identity statements.

6.5. Objections to DLB

6.5.1. Objection 1

As a way of introducing the first worry, recall that by assumption Max has the names 'Cicero' and 'Tully' in his idiolect, which is to say that he believes the Assignments given by (24) and (32), whose logical forms are again those in (24') and (32'), respectively:

(24) 'Cicero' refers to Cicero

(32) 'Tully' refers to Tully

(24') '$[\text{Cicero}]_1$' refers to $[\text{Cicero}]_1$

(32') '$[\text{Tully}]_2$' refers to $[\text{Tully}]_2$

Now, as noted repeatedly, coreference of '[Cicero]$_1$' and '[Tully]$_2$' is not fixed by the grammar. Yet by assumption these two expressions do in fact corefer, say as determined by their causal-historical chains of reference.[35] If correct, the worry turns on how we as theorists are to understand the formal analysis of these Assignments. If we take them as equivalent to traditional meaning axioms, then the expressions '[Cicero]$_1$' and '[Tully]$_2$' on the right hand side of the "refers to" relation contribute the same individual (i.e., Marcus Tullius Cicero) to the propositions they encode, despite what Max may or may not assent to. If so, then the logical forms of (1") and (2") are perhaps more perspicuously represented as (1'") and (2'"):

(1'") g('[Cicero]$_1$') = Marcus Tullius Cicero

(2'") g('[Tully]$_2$') = Marcus Tullius Cicero

Substituting equals for equals, we should then be able to rewrite F&M's *de dicto* logical forms of, say, (4) and (5) as (4") and (5"):

(4") [Cicero was a Roman statesman]
& ['[Cicero]$_1$' refers to Marcus Tullius Cicero]

(5") [Tully was a Roman statesman]
& ['[Tully]$_2$' refers to Marcus Tullius Cicero]

The obvious question is of course if Max believes each of (1'"), (2'"), and (4) then what blocks his inference to the truth of (5)? For understood this way, the example above appears to be an instance of the familiar inference scheme: $(Fa \ \& \ a = b) \supset Fb$. In other words, F&M's proposal appears to suffer from precisely the same defect that it was designed to solve.[36]

[35] F&M seem to accept some such story. I doubt that this is true, but that doesn't matter for purposes here.
[36] This objection owes to Gary's Ostertag's (2007) review of *DLB*. I understand that David Braun raised a similar objection during a panel discussion of *DLB* at the APA Pacific conference in 2009.

6.5.2. Objection 2

A second serious worry, raised in Ostertag (2007), has to do with so-called "negative" belief reports, as in (40),

(40) Max does not believe that Tully was a Roman statesman

whose *de dicto* logical form is represented by (40"):

(40") Max does not believe that [[Tully]$_2$ was a Roman statesman] & ['[Tully]$_2$' refers to [Tully]$_2$]

As Ostertag rightly notes, (40") is compatible with Max not believing the 'Tully'-Assignment that is justifiably attributable to him by the *de dicto* logical form of (12) given in (12"):

(12) Max believes that Tully was a Roman poet

(12") Max believes that [[Tully]$_2$ was a Roman poet] & ['[Tully]$_2$' refers to [Tully]$_2$]

This in turn suggests that F&M's conjunctive analysis of *de dicto* logical forms, as discussed earlier, is seriously flawed. In Ostertag's words:

> If a *de dicto* report attributes belief in *P & A* (where *A* is an assignment) then contradicting that report should simply consist in a denial of belief in *P & A*. But whereas one can use a negative *de dicto* belief report to deny belief in *P*, one cannot use a negative *de dicto* belief report to deny belief in *A* (to do so is incompatible with the report's being *de dicto*). This suggests that the conjunctive analysis is incorrect—that *A* is not part of what is asserted.

Ostertag cites other worries for a metalinguistic approach to the puzzles, including those raised by Saul (1997), though I won't rehearse those details here. The alternative psychological construal of F&M's solution as proposed in the beginning, however, arguably avoids these formal semantic issues.

6.6. Chapter summary

The preceding has been a particularly long-winded way of getting at the same worry that Kripke (1979) raises about the defective nature of our belief-reporting practices, or at least under the assumption that the meaning/propositions expressed by belief-reports are supposed to mirror the logical forms of the beliefs they are used to report. For indeed under this assumption the semantics of belief-reports seem ill-equipped to avoid the pitfalls raised in *A Puzzle about Belief*. However, on my view one can explain the *informativeness* of identity statements without taking a stance about the relationship between the meanings of those statements and the beliefs they are used to report. This is the project of Chapter 7, which accepts many of the intuitions motivated in *DLB* while rejecting the claim that what's informative about identity statements is literally expressed as part of their meaning/semantic content. Rather, on view the solution to informativeness is extralinguistic/pragmatic/inferential in nature yet which nonetheless, as F&M argue, crucially involves beliefs about the meanings of Words. In this way, I contend that my proposal avoids the failure of F&M's linguistic-semantic solution to the puzzles. It is worth stressing here that I will not be offering a theory of beliefs, nor a theory of the semantics of belief reports, *per se*. I will however be using some empirically justified conclusions about the linguistically generated meanings of sentences of the form 'α is β' to explain how their utterances relative to a context can be *informative* in the relevant way.

7. A Metalinguistic Solution to Frege's Puzzle of Informativeness

7.1. Introduction

As advertised, in this final chapter I argue in favor of a broadly metalinguistic solution to Frege's puzzle about the informativeness of true natural language identity statements of the form 'α is β', more or less as the puzzle was characterized in Chapter 6, yet which navigates past the difficulties encountered by Fiengo & May (F&M).[1] However, I again concur with F&M on several key points. In particular, I agree that competent speakers have (at least tacit) beliefs about the syntactic identity (or difference) of the names that appear in these statements, and that these beliefs influence speaker-judgments about the reference, or coreference as the case may be, of name-tokens relative to a context/discourse. I further agree that such beliefs play an important role in the interpretation of identity statements for competent subjects who are otherwise ignorant of the truth of the thoughts they express. However, *contra* F&M my positive proposal is that an utterance of 'α is β' *pragmatically conveys* (but does not *express*) the metalinguistic fact that the names 'α' and 'β' designate the same object (i.e., that they *corefer*). This latter proposition, I contend, is derivable solely from the utterance's linguistic form along with the subject's understanding of the semantic role of proper names in the grammar.

What distinguishes my view from other pragmatic accounts such as Salmon (1986) and Soames (2002) is that on my view the metalinguistic information about coreference pragmatically conveyed by 'α is β' provides subjects with a logical premise needed to

[1] While my proposal extends naturally to questions related to belief reports containing proper names, I do not directly address these questions here, save a very brief discussion near the end regarding how my proposal applies to identity statements of the form 'Paderewski is Paderewski' that involve two occurrences of the same name. I also propose to set aside identity statements flanked on both sides of the identity sign by definite descriptions.

further deduce that there exists at most *one* contextually-salient bearer of both names (or *two* in the case of 'α is *not* β'). It is coming to accept this latter proposition regarding not precisely *who* or *what* is being referred to but *how many* objects/individuals are under discussion—a proposition which was not believed (or positively disbelieved) prior to the utterance—that can rationally compel misinformed subjects to accept the truth of 'α is β'. Importantly, my proposal relies on the assumption that the former thought/proposition regarding how many objects/individuals are under discussion may not be the same thought that speakers intend to express with 'α is β' (yet without taking a firm stance as to what, precisely, this latter thought is). As importantly, I argue that accepting the former thought as true constitutes what Frege considered a "substantive" advance in the interpreter's knowledge, which holds irrespective of *what else* they may or may not know about the relevant object(s) of reference. For even if subjects know nothing else apart from their names, they can readily deduce *how many* objects are the topic of discussion. Moreover, these subjects will have thereby learned that whatever they previously believed (*de re*) about the referent of 'α' is also quite likely true of the referent of 'β', and *vice versa*.[2] Furthermore, they will henceforth know that the truth of 'α is β' is what rationally justifies the intersubstitution of 'α' and 'β', *salva veritate*, in transparent contexts.

In short, the central claim of this chapter is that while one might agree with Frege that the "cognitive value" of an identity statement is determined by the thought that it "contains" (i.e., the proposition that it expresses), I contend that the actual informativeness of those statements *for Frege-subjects* derives chiefly from the logical

[2] I say "quite likely true" as I am willing to grant possible exceptions. For instance, I am willing to grant that it may be true, strictly speaking, that while Superman can fly, Clark Kent cannot.

implications of their linguistic forms.[3] At bottom, the relevant information conveyed by an identity statement is again *not*, in my view, *who* or *what* is being referred to but rather *how many* objects/individuals are under discussion. The full details of my proposal require quite a bit of unpacking. And I will set forth the agenda of this chapter momentarily. However, I wish to distance this discussion somewhat from the previous chapter by stating a few additional background assumptions about the target phenomenon, not all of which are explicitly (or even implicitly) endorsed by Fiengo & May. At the same time, I will attempt to further clarify my approach to resolving the puzzle.

7.1.1. Resetting the discussion

First of all, I take it that Frege's puzzle about the informativeness of identity statements involving type-distinct coreferential names, and correlatively failures of substitution of those names *salva veritate* in opaque contexts, both arise for one of two simple reasons. Either (i) subjects do not know/believe *de dicto* that the names in question corefer (or not), or (ii) they have confused *de re* one individual for two (or *vice versa*). These two facts are intimately related, however, in the sense that ignorance of the one *explains* ignorance of the other.[4] For in the general case, to believe of two names that they do not corefer is *ipso facto* to be committed to the belief that those names have numerically distinct bearers. Conversely, to believe *de re* that there exists two

[3] Similar points are raised in Perry (2003), and elsewhere. However, I am unaware of anyone who has tried to develop such considerations into a full-blown hypothesis.

[4] To be clear at the onset, I assume that suitably sophisticated non-linguistic animals can in principle be Frege-subjects. If correct, then Frege's puzzle is not a puzzle about language, *per se*, but rather a puzzle about the nature of *thought*. This point will be repeated below. But in short, given that the puzzle is typically cast as a linguistic phenomenon, and which is how Frege himself conceived the puzzle, I will address my arguments specifically to language users.

objects/individuals that are the respective bearers of type-distinct names rationally entails the subject's *de dicto* belief that those names do not corefer.

Similarly, there at least two ways out of the confusion. Subjects might either (i) learn, perhaps by being told, that the names in question corefer (or not), or (ii) infer from independent evidence that there exists at most one object/individual that happens to be a (or perhaps *the*) bearer of both names (or in the reverse direction, that there exists multiple bearers of the same name). These latter two facts are also related in the formal sense of providing premises that logically connect one to the other such that a subject who both recognizes and accepts the one fact will be rationally compelled to accept the other (or perhaps *should* be so-compelled, and again *ceteris paribus*).

There is of course a third way of becoming so-informed, or anyway by common assumption, which is to be told by way of a true identity statement of the form 'α is β' that α and β are in fact one and the same object/individual. As raised in the previous chapter, the challenge is to specify which property (or perhaps properties) of such statements renders them informative. For again if one takes a Millian view of names then presumably what speakers *express* with 'α is β' is the trivial/uninformative proposition that the object referred to is self-identical. However, from the perspective of misinformed interpreters what is *understood* by such statements, if taken at face value, *appears to them*, at least initially, to be a logical contradiction. As competent listeners know *a priori* that [what they believe to be] numerically distinct objects cannot at once be numerically identical. Competent subjects also know *a priori* that the only way they could be mistaken about the truth of 'α is β' is if the names in question are in fact being used in reference to the same object/individual. Otherwise, it would simply never occur to them

that what is being expressed is the trivial truth that the common referent of ‘α’ and ‘β’ is self-identical.

Practically speaking, a perhaps more effective way for speakers to achieve the desired result is to utter a more fully articulated sentence such as ‘α is numerically identical to β’, or ‘α and β are numerically identical’, or ‘α is the same [ϕ] as β’, or equivalently ‘α and β is/are the same [ϕ]’, and so forth, where ‘ϕ’ is an optional but often helpful modifier such as ‘object’, ‘person’, or what have you. Better yet, in my estimation, speakers might just as well inform addressees that the names ‘α’ and ‘β’ refer to the same object/individual—i.e., that there is just one contextually-salient bearer of both names. As I think the latter thought is the basically what misinformed understand anyway. But setting this question aside for the moment, the point for now is that there appears to be several ways for speakers to achieve the desired effect of communicating the fact that just one object/individual is under discussion. But thanks in part to Frege and Russell we as theorists are tasked to explain the informativeness of isomorphic translations of the formal language expression ‘a = b’ into corresponding sentences of natural language (e.g., ‘α is β’ in English, ‘α ist β’ in German, and so on). And so far as I can tell, most contemporary researchers, including Fiengo & May, agree that with respect to natural language the informativeness of identity statements of the form ‘α is β’ constitutes the core *explanadum* of Frege’s puzzle.

So constrained, I will just go along with the program and take as my target the relative informativeness of true identity statements of the (abbreviated) form ‘α is β’ (as opposed to ‘α is α’ and ‘β is β’). Now, in most discussions of Frege’s puzzle it is also widely assumed that Frege-subjects are antecedently acquainted with the relevant

object(s) of reference, either by direct perceptual acquaintance or by uniquely identifying definite description (or some set of such descriptions). That is, Frege-subjects are presumed to be capable of thinking about those objects *as such*, if perhaps only in a *de re* sense. And recall from Chapter 3 that if Fodor is right, one cannot think about things as such without having a *concept* of those things. In turn, interpreters who lack concepts of the relevant object(s) are presumably incapable of fully grasping the thought expressed by an identity statement. Moreover, if I understand Frege correctly one cannot grasp the cognitive significance of 'α is β' without fully grasping the thought that it expresses. If true, then Frege-subjects cannot grasp the informativeness of 'α is β' unless they have at least one concept of α/β with which to think about that object *as such*. In my view, however, this latter assumption is misguided and I shall challenge it below.

With stated qualifications in place, the dialectical strategy of this chapter will be as follows. Since my aim is to meet Frege's challenge head-on—i.e., more or less as Frege himself conceived the puzzle, though again as applied specifically to natural language—I begin with a more careful review of the developments in Frege's own thought on the matter, starting with his *Begriffsschrift* view of identity followed by his ultimate rejection of that account in *On Sense and Reference*. Part of the rationale behind this initial exercise is that my own proposal can be thought of as a kind of "psychologized" rendering of Frege's *Begriffsschrift* account of identity statements. I then briefly review certain technological resources developed in earlier chapters, and specifically my conception of the semantics of proper names and how they are understood by ordinary language users. I will then apply these resources to demonstrate how Frege-cases are plausibly generated together with a full account of how, along the lines sketched above,

Frege-subjects manage to grasp the cognitive import of identity statements largely on the basis of their linguistic competence (LSC) alone, which I take to include, more specifically, relevant aspects of their metalinguistic-semantic competence (MSC).

7.2. Frege's *Begriffsschrift* account of identity

7.2.1. Foundations

Frege's partitioning of semantic (or conceptual) content into *sense* and *reference*, as first introduced in *Function and Concept* (1891) and later expounded in *On Sense and Reference* (1892), is widely viewed as a corrective to his earlier *Begriffsschrift* (1879) analysis of identity statements in terms of their capacity to warrant the substitution of formally distinct yet content-identical names, *salva veritate*, in the context of a logical proof.[5, 6] For while the Frege of *Begriffsschrift* had an eye toward semantics, which is to say the objective nature of thought [*Gedanke*] and its relation to truth and reference, his main concern *circa* 1879 was to reinvent logic as a purely formal, gapless, and fully general system of reasoning. More specifically, Frege's aim was to demonstrate how the logical form of thought licenses *valid inference*, which is to say logically justified transitions from one propositional form to the next in the derivation of a proof.[7] Identity statements presented Frege with a special problem, however, which revealed a fundamental tension between the demands of logic and those of semantics—i.e., that

[5] To my mind, this interpretation of Frege has convincing textual support; *cf*., e.g., Heck (2003) for a thorough defense, though see Thau & Caplan (2003) and Bar-Elli (2006) for contrary perspectives.
[6] This corrective action would of course have wider-reaching consequences with respect to substitution not only with respect to identity statements but also within so-called "opaque" contexts such as propositional attitude reports, the latter which I again set aside for purposes here.
[7] Where thoughts are taken to be logical entities that exhibit function-argument structure and which serve as premises and conclusions in logical arguments (as opposed to the sentences used to express those premises and conclusions).

which is relevant to valid patterns of inference as opposed to that which is relevant to questions of truth and reference.[8]

The difficulty for Frege arises from his early acceptance of names (*à la* Mill) as signs of objects; i.e., as logical constants that merely "stand for" [*bedeuten*] their referents. For given a language that permits the introduction of formally distinct names with the same *Bedeutung*, those names should, in consequence of Leibniz's law of self-identity, be everywhere substitutable *salva veritate*.[9] Yet one needs a way to express what Frege considered to be a fundamentally semantic fact in a logically perspicuous manner; for example, in a way that licenses judgments/inferences of the form:[10]

(1) |---- $(Fa \supset Fb)$

For (1) to constitute a valid step in a proof, however, one must first establish, usually by mere stipulation, that the names '*a*' and '*b*', understood as formal symbols, are in fact content-identical and thus everywhere substitutable under identity of content. Yet prior to *Begriffsschrift* the only notion of identity available to Frege and his contemporaries was what is sometimes referred to as *objectual identity*, which expresses the fact that the two arguments *qua objects* flanking the symbol '=' are *numerically identical* (or *self-identical*). That is, it was commonplace at the time to assume that statements of the form '*a = b*' express the *material* fact that *a is identical to b*, or that *a is the same object as b*, or as Frege later puts it that *a and b coincide*. The problem, as Frege saw it, is that under objectual identity a judgment such as (2) in the body of a proof,

[8] For Frege of the *Begriffsschrift*, the relevant judgments were judgments of *form* rather than judgments of *assertable contents*, the latter being defined as the transition from a thought to a truth-value, and which would ultimately take center stage in Frege's later work.

[9] As stated by Leibniz, the law reads: "those things are the same of which one can be substituted for another without loss of truth."

[10] In Frege's notation, the horizontal line is the "content stroke" and the vertical line is the "judgement stroke." I translate Frege's rather cumbersome notation for conditionality into the form $Fa \supset Fb$ for sake of readability.

(2) |---- ($a = b$)

fails to express what is needed to *formally* justify the substitution of '*b*' for '*a*' in statements like (1), thereby rendering the step invalid. As with respect to (1), it is the substitution of *names* and not their objects (or *contents*), that is logically relevant.[11]

Frege's *Begriffsschrift* solution to the problem—what we might call his *metalinguistic analysis of identity*—was to introduce a new symbol, the triple-bar '≡', to designate what he called *identity of content* [*Inhaltsgleichheit*], which is basically what we today think of as formal equivalence.[12] For what this relation does is license the intersubstitution of formally distinct names, *salva veritate*, in the context of proof statements such as (1). For instance, a well-formed formula of Frege's *Begriffsschrift* is given by (3),

(3) |---- ($a \equiv b$)

which can be read as asserting the judgment that "*the symbol a and the symbol b have the same conceptual content, so that a can always be replaced by b and conversely*."[13] Thus, while Frege assigns a special meaning to the symbol for *identity of content* he retains the classical Leibnizian clause regarding the substitution of equals for equals. What is needed in addition, however, is yet another proposition (or perhaps a rule of inference) that

[11] As May (2000: 9) puts it, "while inferential relations hold between thoughts, we can only determine whether one thought follows from another with respect to the *form* by which that thought is expressed."
[12] The German term '*Inhaltsgleichheit*' is sometimes translated as "identity of content." For as Beaney (1997: 64, fn.24) notes, "Throughout his writings, both here [in *Begriffsschrift*] and in his later work, Frege makes clear that he understands '*Gleichheit*' in the sense of 'identity'; so either rendering is correct." This point will be reinforced below.
[13] Frege uses the capital letters 'A' and 'B' but I am taking the liberty of replacing them with lower case letters to maintain consistency with earlier examples. Frege also sometimes ignores the use/mention distinction. But in this case he is clearly speaking about symbols and not their contents.

justifies the transition from (3) to (1). In *Begriffsschrift*, this lacuna is filled by Frege's basic proposition #52, whose equivalent formulation is given in (4):[14]

(4) |---- $(a \equiv b) \supset (Fa \supset Fb)$

In paraphrase, (4) states that one may freely substitute the symbol '*b*' for '*a*' in any formula so long as '*a*' and '*b*' are content-identical; a fact which again must first be established by asserting (3) somewhere in the body of the proof. In short, what we wind up with, by *modus ponens*, is the following sequence of formally valid judgments:

(3) |---- $(a \equiv b)$

(4) |---- $(a \equiv b) \supset (Fa \supset Fb)$

---

(1) |---- $(Fa \supset Fb)$

The chief drawback of this proposal is that not only does it bifurcate the logician's construal of identity as between *objectual identity* and *identity of content*—a situation that would ultimately not sit well with Frege—it also bifurcates the meaning of each atomic name in the object language. For in all other contexts names again merely stand proxy for their referents. However, in the context of a statement of identity of content names come to stand for themselves. Or as Frege puts it in his opening remarks of §8 of *Begriffsschrift*:

> Equality [or Identity] of content [*Inhaltsgleichheit*] differs from conditionality and negation by relating to names, not to contents. Elsewhere, signs are mere proxies for their content, and thus any phrase they occur in just expresses a relation between various contents; but names at once appear *in propria persona* so soon as they are joined together by the symbol for equality [/identity] of content; for this signifies the circumstances of two names having the same content. Thus, along with the introduction of a symbol for

[14] As reported by May (2000), in "Boole's logical Calculus and the Concept-script," published shortly after *Begriffsschrift* in 1881, Frege considers exporting his basic proposition #52 as a rule of inference, which leads to certain complications that are largely irrelevant to my purposes here.

> equality [/identity] of content, all symbols are necessarily given a double meaning—the same symbols stand now for their own content, now for themselves.

Frege's *Begriffsschrift* notion of *identity of content* is therefore a fundamentally *metalinguistic* construal of identity, as statements of the form '$a = b$' express a relation between the names '*a*' and '*b*'. In particular, as Frege puts it above identity expresses the circumstance that two names have the same content.

7.2.2. Justifying coreferential terms

To avoid this ambiguity, one might wonder why Frege did not instead simply prohibit the introduction of content-identical names into his logical language. For if the relation between names and their objects were strictly one-to-one there would be no need for statements of identity of content and hence no need for a special symbol to express this relation. However, Frege realized that such a constraint would come at the cost of expressiveness. Specifically, it would sacrifice the full generality of his *Concept-script*, which was paramount to the success of his wider logicist program whose ultimate goal was to reduce arithmetic to logic. As without the ability to introduce content-identical names, along with perhaps some "fruitful" definitions,[15] one cannot derive certain proofs/theorems that might otherwise be derivable with their aid.

To motivate this exception to the rule, Frege invokes the geometric example in Figure 1 to demonstrate the need for formally distinct names with the same content:

[15] *Cf.*, Horty (2007) for a thorough and insightful analysis of Fregean definitions.

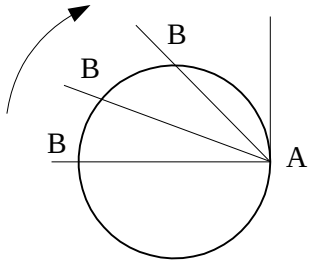

Figure 1. *As line B representing the circumference of the circle turns about point A in the direction of the arrow, line B moves toward point A until they coincide at point A.*[16]

Frege observes that where the perpendicular line intersects the circle at the point named 'A', the name 'B' has the same content as 'A', yet whose shared content is determined in two different ways. Specifically, he argues (again from §8 of *Begriffsschrift*):[17]

> The need for a symbol for identity of content thus rests on the following: the same content can be fully determined in different ways; but that, in a particular case, *the same content* is actually given by *two modes of determination* is the content of a *judgement*. Before this judgement can be made, two different names corresponding to the two modes of determination must be provided for that that [the relevant content] is thereby determined. But the judgement requires for its expression a symbol for identity of content to combine the two names. It follows from this that different names for the same content are not always merely a trivial matter of formulation, but touch the very heart of the matter if they are connected with different modes of determination.

That two names might differ in their "modes of determination" [*Bestimmungsweise*], or what I will call MoDs for short, is signaled by their distinct *forms* (or *shapes* in the case

[16] This figure is egregiously reproduced (though not copied) without permission from Beaney (1997: 64).
[17] The translation here, taken from Beaney (1997), is I think more comprehensible than the Geach & Black (1960) translation, yet which I cite for comparison as follows:

"The need of a symbol for equality of content thus rests on the following fact: The same content can be fully determined in different ways; and *that*, in a particular case, *the same* content actually is given by *two ways of determining it*, is the content of a *judgement*. Before this judgement is made, we must supply, corresponding to the two way of determination, two different names for the thing thus determined. The judgement needs to be expressed by means of a symbol for equality of content, joining the two names together. It is clear from this that different names for the same content are not always just a trivial matter of formulation; if they go along with different ways of determining the content, they are relevant to the essential nature of the case."

of atomic symbols such as syntactically simple proper names). Normally a difference in form signals a *difference* in content. But in the present example the fact that two formally distinct symbols are content-identical must again be established somewhere in the body of the proof and is why Frege needed a custom symbol to express this relation.

Indeed, it is not entirely clear from the quotation above precisely what work MoDs are supposed to be doing for Frege. Yet the following passages from May (2000) are suggestive of an answer:

> (17) What Frege has illustrated here is a "proof" of the judgement that "A" and "B" have the same content; the role of the modes of determination is that they stand as premisses of this proof. These modes of determination, however, are *synthetic*, one being geometrical, the other perceptual. Thus, because of the way in which this conclusion was reached, in this case, Frege puts it, "the judgement as to identity of content is, in Kant's sense, synthetic."

And earlier May explains:

> (15-6) Modes of determination, Frege thus tells us, are what *justify* judgements of identity of content; but more than that, it is through this justification that we can touch the thought, for this justification will provide information fixing the judgements within the [traditional Kantian] categories of thought.

On May's interpretation, in other words, Frege is employing MoDs as a formal way to distinguish statements of identity of content as either *a priori analytic* versus *a posteriori synthetic*. Distinct MoDs in the context of an identity statement therefore signal what *kind* of thought one is faced with. And as justified in this way judgments of identity of content make contact with the proper subject matter of logic—i.e., the judgeable contents of thought—albeit only indirectly, as May qualifies:

> (16) The justification to which Frege refers, we must understand, is not itself part of the content of the judgement per se, and is thus not represented in the judgement; it is rather extra-notational. Modes of determination are thus in no way represented in judgement

> above and beyond that which follows about them in virtue of the occurrence of (distinct) symbols. What is *expressed* by a judgement of identity of content via its representation in the conceptual notation is solely that the symbols have identical content; the justification for such a judgement, and hence its connection to "thought," is not so represented in such logical forms.

Or as Frege puts in the *Grundlagen* (Beaney [1997: 92]):

> Thus in general the question as to how we arrive at the content of a judgment has to be distinguished from the question as to how we provide the justification for our assertion.

What specifically follows from an identity of the form 'A ≡ A' (or '*a* ≡ *a*') is that it is an instance of Leibniz's law and thus trivially (i.e., self-evidently or necessarily) true, and indeed known to be so *a priori*; that is, regardless of what 'A' (or '*a*') refers to. Put differently, we can say that a logician is justified in judging such a thought true on the basis of its form alone.

In contrast, what follows immediately from the linguistic form of 'A ≡ B' (or '*a* ≡ *b*') is that the thought contained therein is *synthetic* and thus knowable only *a posteriori*. May elaborates as follows:

> (17) Judged by conceptual content, introducing a new atomic symbol into the conceptual notation with the same content as some other atomic symbol, does not, in and of itself, change the expressiveness of the system, unless by that introduction there is some judgement that can be expressed that otherwise could not be. This obtains in the case at hand, for a synthetic judgement has become possible, established by a proof of that judgement, where the critical premisses in the proof that fix its status are the modes of determination. Without such modes of determination there would be no way to establish that distinct expressions, atomic or not, play distinct roles in proofs, even though they have the same content.

Notice that while MoDs resemble what Frege would later call "modes of presentation," they are not the same notion. For as May suggests, the MoDs of names are not for Frege

part of the content expressed by sentences in which those names occur. However, there again seems to be a sense in which MoDs are at least indirectly related to their contents, for as May further clarifies:

> (18) Such assertions are not just about the expressions, but also pertain to the thought because each atomic term is itself justified by being associated with distinct modes of determination. Such distinct modes of determination thus give good reason for multiple atomic terms for a given object… (21) It is characteristic of [judgments of identity of content] that significant information relevant to conceptual content is secreted away, revealed as modes of determination… Where modes of determination are called for is just where there is a relation [identity of content] that does not reveal the relevant information, in particular where we say *of two atomic names* [my emphasis] that they have the same conceptual content.

In other words, it is important to stress that MoDs are *not* by Frege's lights part of the judgeable content represented by (3) above but rather serve as the formal justification for making that judgment. For again what is *expressed* by (3), unlike (2), is the *metalinguistic fact* that the symbols '*a*' and '*b*' are content-identical. In turn, to recognize an identity statement as involving type-distinct names is to recognize those names as being associated with distinct MoDs. Given two different MoDs, one can thereby "see" *a priori* that (3) is not merely an instance of Leibniz's law (e.g., '*a* = *a*') but rather that the thought contained therein is synthetic and thus carries substantive information about its proper subject matter, which is to say information about its object of reference.

Of course one wants to know what, exactly, this substantive piece of information amounts to; that is, beyond the metalinguistic fact that the names '*a*' and '*b*' are content-identical? It can't be, as Russell seemed to have thought, that their common referent is self-identical. For as Frege later acknowledges in *On Sense and Reference*, any competent logician who feels justified in asserting '*a* = *b*' presumably knows *a priori* that

the object being referred to is self-identical. He also knows by mere inspection that '*a*' and '*b*' have distinct MoDs. But this latter fact says nothing about the actual identity of their shared content in the sense of uniquely identifying/describing precisely *which object* (i.e., who or what) is being referred to. Yet if one is logically and/or rationally justified in asserting that the names '*a*' and '*b*' are logically equivalent one can nevertheless justifiably infer, *a priori, how many* objects are being referred to.

Let me pause here to stress that the purpose of this discussion is to tease out what, on Frege's view, a competent logician can justifiably infer from the purely formal properties of an identity statement of the form '*a* = *b*'. For purposes of conducting a formal proof, what is logically relevant is having proved that '*a*' and '*b*' are content-identical. And from this one can be sure that whatever is true of the referent of '*a*' is also true of the referent of '*b*', which is again the feature of the proof that licenses their intersubstitution. With respect to natural language, I contend that what ordinary language users are justified in inferring from the presumed truth of an utterance of 'α is β' is that there is just one object/individual so-called. Moreover, as agued in greater detail below, knowing this much can constitute a substantive advance in knowledge, particularly if such knowledge licenses other material inferences about the relevant object(s) of reference.

7.2.3. A note of historical interest

By way of contrast, it's worth briefly mentioning that by the publication of *Grundlagen der Arithmetik* (1884) Frege appears to have set aside his *Begriffsschrift* account of identity in favor of what we might call his *objectual* (or *semantic*) *analysis of identity* expressed by the traditional mathematical symbol of equality ('='). The reason

appears to be this. First, recall that Frege's goal in the *Grundlagen* is to analyze the concept of number in purely logical terms. Unlike the possibly many-to-one relationship between logical symbols and their contents, however, the relationship between numerals and numbers is always one-to-one as fixed by the grammar of arithmetic. Even in cases where a complex expression such as '2 + 4' appears to designate the same number as '6', there is no need for a statement of identity of content such as '2 + 4 ≡ 6' to convey the information that the signs '2 + 4' and '6' have the same content—i.e., that they designate the same number in two different ways. For the expression '2 + 4' is not for Frege a complex *name* for the number six but rather a complex expression/formula that contains the names of two numbers, *viz*., two and four, whose sum equals six.

Unlike logical identities, in other words, judgments of mathematical equalities such as '2 + 4 = 6' are justified by information that stands in plain view of the mathematician. Hence, there is no need to convey information (e.g., MoDs) beyond what is given by the content of such statements. What's more, since Frege believed that all mathematical statements are at bottom *analytic*, there is no need for different MoDs, as exists in logic and geometry, to distinguish analytic statements from synthetic ones.

Indeed, in the *Grundlagen* Frege *seems* to have abandoned his *Begriffsschrift* notion of identity of content once and for all in favor of a unified conception of *Gleichheit* as what I am calling *objectual identity*. He writes in §63, for example, that:

> The relationship of equality does not hold only amongst numbers. From this it seems to follow that it ought not be defined specially for this case.

Frege then proceeds to demonstrate that even the geometrical judgment that two lines are parallel, traditionally indicated by the symbol '//', can be specified in terms of equality

('=') whereby two parallel lines are "defined as lines whose directions are equal" (§64). In short, May writes (*ibid*: 25):

> As Frege now sees matters, with a rather revisionist ring given his own prior view, the notion that mathematicians (including himself) actually had in mind when they employed the equality symbol was identity, and it is this "ordinary sign" that subsumes identity of content. But although like equality it stands for an objectual relation, it is not a symbol peculiar to mathematics, as is equality; rather it is a general symbol, applicable to propositions about all sorts of things, as is identity of content.

Yet it is not entirely clear whether prior to *On Sense and Reference* Frege has actually abandoned his notion of identity of content altogether. For his only justification for treating mathematical equality as being on par with objectual identity is that the former applies to a domain less general than logic. Rather, on my reading Frege finds himself stuck with two distinct notions of identity, neither of which is general enough to apply to across the board. And this, I take it, is why we find Frege still wrestling with the question in 1892.

All this being so, let me fast-forward to a review of Frege's opening passage of *On Sense and Reference* (or OSR for short) where according to standard interpretation Frege attempts to overturn his previous two conceptions of identity in favor of a third, which follows from, or perhaps rather justifies, his distinction between the *sense* of a linguistic expression and its *reference*.

7.3. On the distinction between Sense and Reference

7.3.1. The opening paragraph

Frege begins his discussion in OSR by again revisiting the problem posed by '*Gleichheit*', which is often translated into English as *equality*:[18]

> Equality [*Gleichheit*] gives rise to challenging questions which are not altogether easy to answer. Is it a relation? A relation between objects, or between names or signs of objects?

In an immediate footnote, however, Frege qualifies that he takes the term '*Gleichheit*' to be synonymous with '*Identität*', the latter which straightforwardly translates to English as *identity*. In this note Frege also appears to advertise his new and improved construal of identity:

> I use this word [i.e., '*Gleichheit*'] in the sense of identity [*Identität*] and understand '*a* = *b*' to have the sense [*dem Sinne*] of '*a* is the same as *b*' ['*a ist dasselbe wie b*'] or '*a* and *b* coincide' ['*a* und *b fallen zusammen*'].

What Frege seems to imply here is that the sense of the expression '*a* = *b*' is synonymous with the sense of '*a* is the same as *b*' (or that of '*a* and *b* coincide'). Or in any case, it would seem that to understand the sense of the former just is to understand the sense of the latter.

What remains unexplained at this preliminary stage, however, is what exactly the names '*a*' and '*b*' contribute to the compositionally determined sense of '*a* = *b*'. I will say more about this below, but Frege ultimately suggests that the sense of a declarative sentence—or again in neo-Fregean terms its *meaning*—can be understood in terms of the *thought* that it expresses. And since on Frege's view the sense of a name determines its reference, the thought expressed by '*a* = *b*', and correlatively '*a* is the same as *b*' (or '*a*

[18] See, for example, Geach & Black (1960) and Beaney (1997).

and *b* coincide'), is the fact that the respective senses of the names '*a*' and '*b*' *determine the same reference*, which is to say that they present the same object in two different ways. In purely epistemic terms, it would therefore seem that to grasp the thought expressed by an identity statement depends on grasping the respective senses of the names involved. In turn, grasping the sense of a name seems to require knowing *how* to determine its reference. However, grasping the respective senses of '*a*' and '*b*' in no way guarantees that subjects will *recognize* that they do in fact determine the same reference, which is presumably how, on Frege's view, Frege-cases are generated.

Frege devotes the remainder of this paragraph to explaining what identity *cannot be*, and in particular to cast doubt on his previous two conceptions of identity, beginning with his *Begriffsschrift* account. Continuing from the first quotation above, he adds:

> In my *Begriffsschrift* I assumed the latter [i.e., that identity is a relation "between names or signs of objects"]. The reasons which seem to favour this are the following: $a = a$ and $a = b$ are obviously statements of differing cognitive value [*Erkenntniswert*]; $a = a$ holds *a priori* and, according to Kant, is to be labeled analytic, which statements of the form $a = b$ often contain very valuable extensions of our knowledge [*Erweiterungen unserer Erkenntnis*] and cannot always be established *a priori*…

As suggested, Frege correlates the difference in *cognitive value* between '$a = a$' and '$a = b$' with a difference in their respective classification among the Kantian categories as either *analytic* or *synthetic*, and as being knowable *a priori* versus *a posteriori*. From a semantic perspective, Frege again considered statements of the form '$a = a$' to be analytic, which is to say true in virtue of meaning alone. Epistemically, the truth of '$a = a$' is knowable *a priori* because it is recognizable as an instance of Leibniz's law solely on the basis of its linguistic form; i.e., without regard to what the name '*a*' denotes. By contrast, statements of the form '$a = b$' are classified as *synthetic* because their truth is

determined by applying a Fregean *Concept* (i.e., *function*) to its argument(s). And only once synthesized can the resulting thought be judged either true or false. If it turns out that what the predicate of the sentence expresses is not, as Kant would say, already "contained" in its subject, then it expresses a synthetic truth that is knowable only *a posteriori*. And for this reason to grasp the content of '*a* = *b*' can constitute a substantive advance in the interpreter's knowledge of the mind and language-external world.

In general, Frege seems to have held that all true *synthetic* statements have cognitive significance, including not only identities of the form '*a* = *b*' but also predicational and relational sentences of the form '*Fa*' and '*Rab*', respectively. For instance, an utterance of '*Fa*' can clearly be informative for a subject who does not antecedently believe the thought that it expresses. Indeed, to merely judge (correctly) that thought *as synthetic* is to recognize it as containing a potentially valuable piece of knowledge. And while Frege is not explicit about this, it appears that just as competent speakers can know *a priori* that '*a* = *a*' is analytic and thus trivially true, so too can one know *a priori* that '*a* = *b*' is synthetic and thus at least potentially informative simply by recognizing that it is *not* an instance of Leibniz's law. To put the point differently, subjects can know whether or not a given identity is informative based solely on what Perry (2003: 9) calls its "epistemic profile."

Crucially, however, it is again the *thought* expressed by a statement that according to Frege *determines* its cognitive value, and in turn its status as either analytic or synthetic. And while it may be true that all synthetic statements are potentially informative, it would seem that statements expressing different thoughts differ in

cognitive value.[19] The relevant point is that it is one thing to recognize *that* a given statement contains a potentially valuable piece of information and quite another to know precisely *which* piece of information is contained therein. However, as I argue in greater detail below, one can nonetheless learn something substantive from a true identity statement of the form 'α is β' simply by recognizing its status *as expressing a non-trivial thought*. Namely, one can infer that there is just one contextually-salient bearer of the names 'α' and 'β', and thus one object/individual under discussion.

Moving along, Frege next observes the following:

> Now if we were to regard equality as a relation between that which the names '*a*' and '*b*' designate [*bedeuten*], it would seem that *a* = *b* could not differ from *a* = *a*; i.e. provided *a* = *b* is true. A relation would thereby be expressed of a thing to itself, and indeed one in which each thing stands to itself but to no other thing.

Frege is here apparently refuting what I earlier characterized as his *objectual analysis of identity*. As mentioned in the previous chapter, the idea seems to be that since everything is self-identical, and competent subjects knows this *a priori*, then every statement of identity between a thing and itself is *analytically* true, hence thoroughly *uninformative*, and indeed recognizable as such *a priori*. Yet in the case of '*a* = *b*' this is clearly *not* the case. Frege therefore concludes that a purely objectual analysis of identity cannot be correct.

Frege then proceeds to reevaluate the status of his *Begriffsschrift* account of identity as *identity of content*, which I have also characterized as his *logical analysis of identity*, or what has now become known as his *metalinguistic analysis of identity*:

[19] Frege is actually a bit unclear about this as well, but it raises a potentially interesting question. For if all that Frege meant is that the thought expressed determines its epistemic profile then subjects could grasp the cognitive value simply by recognizing its epistemic profile. However, I seriously doubt that this is what Frege intended, and I assume not in what follows.

> What we apparently want to state by *a* = *b* is that the signs or names '*a*' and '*b*' designate the same thing, so that those signs themselves would be under discussion; a relation between them would be asserted. But this relation would hold between the names or signs only in so far as they named or designated something. It would be mediated by the connection of each of the two signs with the same designated thing. But this is arbitrary. Nobody can be forbidden to use an arbitrarily producible event or objects as a sign for something. In that case the sentence *a* = *b* would no longer be concerned with the subject matter, but only with its mode of designation; we would express no proper knowledge [*eigentliche Erkenntnis*] by its means.

In other words, Frege again understood all too well that the assignment of names to individuals is *arbitrary*, which is to say a matter of linguistic convention. It is then equally arbitrary whether any two names designate the same individual. And so to assert '*a* = *b*' is to merely report a *metalinguistic* fact about language *use* as opposed to a *substantive* (i.e., *material*) semantic fact about the relevant object of reference. And so in this sense statements of the form '*a* = *b*' are no more informative than statements of the form '*a* = *a*', as we would again "express no proper knowledge by its means." Yet Frege continues:

> But in many cases this is just what we want to do. If the sign '*a*' is distinguished from the sign '*b*' only as an object (here, by means of its shape), not as a sign (i.e. not by the manner in which it designates something), the cognitive value of *a* = *a* becomes essentially equal to that of *a* = *b*, provided *a* = *b* is true. A difference [in cognitive value] can arise only if the difference between the signs corresponds to a difference in the mode of presentation of the thing designated.

Thus, we have arrived at the final stage in Frege's mature thought on the matter, which is that names '*a*' and '*b*' have different *modes of presentation* [*Art des Gegebenseins*].

More specifically, we find Frege exploiting the perceived difference in cognitive value between '*a* = *a*' and '*a* = *b*' as way of motivating his semantic distinction between the *sense* of an expression and its *reference*. Or put the other way around, Frege's sense-

reference distinction is invoked to *explain* the presumed difference in cognitive value between '*a* = *a*' and '*a* = *b*'; that is, to account for why the latter is informative in a way that the former is not. For Frege reasoned that since these two statements obviously differ in cognitive value they must express different thoughts.[20] And the reason they express different thoughts is because the names '*a*' and '*b*' have different *senses*, which again in neo-Fregean terms is to say that they differ in *meaning*.[21] As such, '*a*' and '*b*' make different semantic contributions to the thoughts expressed by sentences that contain them. This in turn explains why their substitution in otherwise structurally identical sentences yields a different thought, and indeed can change the *kind* of thought expressed by such statements, generating a synthetic statement from one that is analytic. And for these reasons, so far as I can tell, Frege concludes that the thought expressed by a declarative sentence is what determines its cognitive value.

7.3.2. Summing up

Summarizing, Frege held that the thought expressed by an identity statement, or for that matter any declarative statement, is classifiable as either analytic or synthetic. Thoughts that are classified as synthetic contain what Frege called "real" cognitive value, and those that are analytic are trivially true and thus thoroughly uninformative. Among those that are synthetic, different thoughts presumably carry different cognitive value in that they contain different information about their object(s) of reference. Hence, as I understand Frege two sentences can be informative yet differ in cognitive value. For example, both '*a* = *b*' and *Fa* (assuming them true) are potentially informative for a

[20] However, as often pointed out in the secondary literature Frege does not bother to defend this assumption as he takes it to be self-evident.
[21] Or what the Frege of *Begriffsschrift* referred to as "conceptual" or "judgeable" content.

subject who is ignorant of the respective truths they express. Yet because they express different thoughts they presumably differ in cognitive value.

Epistemically speaking, our *judgments* about the truth or falsity of a given statement are either *a priori* or *a posteriori*. A subject who correctly judges a statement to be analytic is entitled to judge it true *a priori*, whereas statements correctly judged to be synthetic require a further *a posteriori* (i.e., empirical) judgment regarding their truth or falsity. In addition, I have suggested on Frege's behalf that with respect to identity statements any competent speaker will pretty much know *a priori* whether the thought expressed by such statements is analytic or synthetic on the basis of their linguistic form alone. Specifically, any identity statement that is not *a priori* recognizable as an instance of Leibniz's law is knowable *a priori* to express a synthetic thought that contains a potentially valuable piece of information. Curiously, Frege never returns to tells us precisely *which* thought an identity statement of the form '*a* = *b*' expresses. But recall from the footnote above that he understands '*a* = *b*' to have the same sense as '*a* is the same as *b*' (or '*a* and *b* coincide'). And given what he says elsewhere I think we can safely take Frege as holding that the thought expressed by '*a* = *b*' is the fact that the respective senses of the names '*a*' and 'b' determine the same reference, or again that they present the same object in two different ways.

To say that senses of the names '*a*' and 'b' determine the same reference could be taken to express the thought that there is just one object/individual that is referent of both names. For as Frege allows, from a psychological standpoint one might grasp the sense of a name without thereby knowing which object/individual it refers to. Moreover, Frege allows that subjects who grasp the respective senses of '*a*' and '*b*' need not thereby

recognize that they determine the same reference. Indeed, Frege-cases presumably arise because Frege-subjects associate the names '*a*' and '*b*' with different senses (or modes of presentation). Or as a neo-Fregean would put it, Frege-cases arise because subjects assign to these names different meanings. I am willing to grant that this latter fact plays an important role in the interpretation of identity statements, yet without being committed to the view that meanings are Fregean senses. Rather, on my view, which is again more or less Chomsky's, linguistic meanings are *intrinsic* properties of expression-types that are understood by interpreters as instructions to activate semantically related extralinguistic concepts. On this way of thinking, to evaluate how ordinary language users interpret natural language identity statements of the form 'α is β' we must evaluate how interpreters understand their lexical constituents from a purely linguistic standpoint.

In the next section, I briefly recapitulate my assumptions about the semantics of proper names as developed in Chapters 1 and 2. I will then rehearse a fairly common analysis of how Frege-cases are generated, conceptually speaking. Once these preliminaries are complete, I will present my proposed solution to Frege's puzzle of informativeness as cast in these terms. To repeat an earlier advertisement, I think of my proposal as essentially a psychologized rendering of Frege's *Begriffsschrift* account of identity as discussed above, except again as applied to ordinary speakers of natural languages. Specifically, I argue that the informativeness of true identity statements of the form 'α is β' is largely determined by the logical implications of their linguistic forms—i.e., the set of rational inferences that their presumed truth supports. As such, identity statements can be informative for subjects who are not antecedently *en rapport* with their relevant objects of reference.

7.4. A brief review of the semantics of proper names

Recall from Chapter 2 that I am taking proper names, like other Words, to be Chomskyan *lexical items*, which is to say phonological forms paired with their linguistically-specified meanings. More specifically, and in line with Fiengo & May (2006), I have characterized lexical items as unrepeated (non-homophonous), syntactically primitive morphological constituents of the grammar that are statically recorded in a speaker's lexicon and which appear as terminal nodes in syntactic trees. With respect to proper names, in particular, I have argued alongside Burge (1973) and others that [the meanings of] names are naturally understood as ordinary one-place predicates of things so-called.[22] When a name is used predicatively, for example, I assume that it is naturally understood to express what we might think of as its context-invariant *meaning*, or in conceptual terms what I have described as the fundamentally metalinguistic concept$_W$ IS-CALLED(PN(w)) whose content can be characterized as the property of *being called by the name 'PN'*. By hypothesis, the conceptual constituent PN(w) is a concept$_W$ *of the name* itself whose content I have characterized as *being the name 'PN'* (as that name is encoded in its speaker's lexicon).

As stressed in earlier chapters, this predicative/metalinguistic construal of the semantics of proper names is not to deny that names can be and indeed typically are used/uttered in reference to specific objects/individuals. Specifically, I have again argued with Fiengo & May and others that names optionally combine, syntactically, with a referential index to form non-terminal *lexical expressions* (LEs) that contain those names as proper constituents. In these contexts the compositionally determined semantic value

[22] As also mentioned in a footnote in Chapter 5, I bracket the expression [the meaning of] because technically speaking it is redundant. For as stated back in Chapter 1 I assume that to understand an expression just is to understand its meaning.

of an indexed LE relative to a context/discourse will be a specific object/individual. For example, when used referentially the name 'Hesperus' will occur as a constituent of an indexed LE such as (5) below, whose compositionally determined semantic value is specified in (6), where again *g* is an interpretation function that assigns values to Tarskian variables (i.e., indices) relative to a sequence, σ.

(5) $[_{NP}\ [_{PN}$ Hesperus$][x_i]]$

(6) $g([_{NP}\ [_{PN}$ Hesperus$][x_i]])^{\sigma} = \lambda x$: x satisfies 'Hesperus' iff $x = \sigma(i)$ & $\sigma(i)$ is called 'Hesperus'

According to (6) the semantic value of the LE in (5) relative to any sequence, σ, of entities in the domain is the i-th object in σ, for $i > 0$. The category label 'PN' again ensures that name tokens apply only to entities that are in fact so called. For instance, the label 'PN' formally restricts the satisfaction conditions of the nominal LE in (5/6) to all and only those objects in the domain called 'Hesperus'.

It is important to remember, however, that in order to interpret (5) one must first (or also) interpret the *bare* (i.e., *non-indexed*) nominal expression '$[_{NP}\ [_{PN}$ Hesperus$]]$', which again by hypothesis expresses, and is understood to express, the concept_W IS-CALLED(HESPERUS(w)), or what we might think of as simply its meaning. In turn, we can think of (5) itself as expressing the concept_W IS-CALLED(HESPERUS, HESPERUS(w)), which I have construed as a particular instantiation of the "doubly unsaturated" concept IS-CALLED(*C*, *C*) where the first argument position is saturated by a contextually-salient singular concept of objects/individuals, and the second by a concept_W of the name itself, which in the present example is to say HESPERUS(w).[23] In the reverse direction, a

[23] I again assume that this configuration is largely determined by a listener's recognition of the speaker's referential intentions.

referential use of the name ‘Hesperus’ will be *understood* as an instruction to activate its interpreter’s concept IS-CALLED(HESPERUS, HESPERUS(w)), which *ipso facto* involves activation of its two conceptual constituents HESPERUS and HESPERUS(w)). Thus, in short, a referential interpretation of the name ‘Hesperus’ will normally result in activation of all *three* concepts; i.e., HESPERUS, HESPERUS(w), and IS-CALLED(*C*, *C*), where *C*= HESPERUS and *C*= HESPERUS(w).

In addition, I assume that activation of these concepts will normally be accompanied by the activation (or priming) of certain contextually salient *beliefs* that contain these concepts as proper constituents. For instance, coming to think about Hesperus with HESPERUS in consequences of having interpreted the name ‘Hesperus’ referentially will naturally prime certain contextually-salient beliefs about Hesperus (e.g., that Hesperus is a planet, that it shines brightly in the evening sky, and so forth). In turn, activation of HESPERUS(w) will typically prime metalinguistic beliefs about what the name ‘Hesperus’ means in terms of which concept (or concepts) that it is customarily used to express. For instance, since I customarily use ‘Hesperus’ to express HESPERUS, my explicit metalinguistic belief about what the name ‘Hesperus’ means is by hypothesis internally represented as something like (7):

(7) ∃w [HESPERUS(w) & EXPRESSES(w, HESPERUS)]

But more accurately, we can say that the content of (7) reflects my tacit understanding of one way in particular that the linguistically-encoded meaning of the name ‘Hesperus’ in my idiolect constrains without fully determining which extralinguistic concepts that name can (and cannot) be coherently used to express in ordinary discourse.

As this concerns the meaning/interpretation of identity statements, what I have just described is what I take to be an empirically motivated account of how ordinary speakers understand [the meanings of] the names that occur in such statements. As previously advertised, in the next section I briefly examine how Frege-cases are generated. Afterwards, I will then turn to a full defense of my proposed solution to the puzzle regarding the informativeness of natural language identity statements of the form 'α is β'.

7.5. Generating Frege-cases

To demonstrate how Frege-cases plausibly arise, yet without endorsing Frege's distinction between sense and reference, I basically follow the diagnosis of Fodor (2008), which is more less the view advocated by Larson & Segal (1995), Strawson (1974), and certain others.[24] As a kind of toy example, consider Hammurabi who together with other ancient Babylonians reportedly believed that Hesperus was the brightest star in the evening sky and that Phosphorous was the brightest star in the morning sky. As the story goes, however, Hammurabi and the others failed to recognize, as did the Greeks before them, that Hesperus and Phosphorous is/are in fact the same object. In consequence, the Ancients also came to believe that they were using the names 'Hesperus' and 'Phosphorous', respectively (or more precisely their Akkadian/Greek equivalents), in reference to two numerically distinct objects.

More specifically, we might assume that as a young amateur astronomer Hammurabi acquired a singular concept of the planet Venus as it appears in the evening sky. Let's again call this HESPERUS, which Hammurabi henceforth used to think about Hesperus *as such*, which is to say the planet Venus as that object appears in the evening

[24] Though I do not recall off the top my head exactly who else defends this kind of view, yet I am certain that there are several others.

sky. Let us further assume that once acquired Hammurabi then used his concept HESPERUS to form the singular belief that Hesperus is the brightest star in the evening sky. Or to borrow a popular metaphor, we might think of Hammurabi's concept HESPERUS as the name of a mental file under which the following set of belief-predicates are stored:

(8) HESPERUS: {STAR(*h*), MAXIMALLY-BRIGHT(*h*), EVENING-VISIBLE(*h*)}

Repeating this sequence of events with respect to Hammurabi's acquisition of PHOSPHOROUS yields the set of beliefs in (9):

(9) PHOSPHOROUS: {STAR(*p*), MAXIMALLY-BRIGHT(*p*), MORNING-VISIBLE(*p*)}

Of course while we as theorists know that *h = p = the planet Venus*, Hammurabi and friends evidently did not. Thus, according to the imagined scenario Hammurabi has acquired two formally distinct concepts, HESPERUS and PHOSPHOROUS, yet without realizing that they are in fact content-identical, which is to say without representing them *as being* content-identical. And this fact, I take it, plausibly explains why subjects in Hammurabi's condition might rationally yet mistakenly believe that that there exists two numerically distinct objects; one that shines brightly in the morning sky and the other that shines brightly in the evening sky.

Importantly, this is not to suggest that Hammurabi failed to know (*de re*) *what* he was thinking about with HESPERUS and PHOSPHOROUS, in the sense of failing to grasp the contents of his own thoughts. Rather, what Hammurabi again failed to recognize is that he was using both concepts to think about the very same object. As such, there is a tolerably clear sense in which Hammurabi was not mistaken in regard to which object he was thinking about with his concepts HESPERUS and PHOSPHOROUS. However, with respect to his thought *processes* HESPERUS and PHOSPHOROUS are by hypothesis formally

distinct mental representations (i.e., mental symbols) that Hammurabi presumably used to generate formally disconnected beliefs about the planet Venus that in turn played different causal-computational roles in his cognition. Specifically, they played different roles with respect to the various *ways* (i.e., *de dicto*) in which Hammurabi thought about the planet Venus. For this reason, Hammurabi could have entertained thoughts (and evidently did) about the same object without thereby being consciously aware of this fact about his inner mental state.

In consequence, one assumes that as a language user Hammurabi also came to believe that the names 'Hesperus' and 'Phosphorous' (or rather again their Akkadian equivalents) have different meanings in the sense of being used to refer to different objects. Although, recall that on my view these names do in fact have different meanings, linguistically speaking, even for us. Specifically, I take it that the *name* 'Hesperus' is understood as an instruction to activate its owner's concept_W IS-CALLED(HESPERUS(w)), whereas 'Phosphorous is understood as an instruction to activate IS-CALLED(PHOSPHOROUS(w)). However, unlike those of us who are in-the-know Hammurabi evidently believed that the names 'Hesperus' and 'Phosphorous' had *mutually exclusive* meanings in the sense of being strictly non-coreferential. And while mistaken, this belief followed naturally from his mistaken belief that there exist two numerically distinct celestial objects; one that shines brightly in the morning sky and the other that shines brightly in the evening sky. Taken together, these facts stand to explain Hammurabi's false belief that the names 'Hesperus' and 'Phosphorous' have numerically distinct bearers.

In short, I again agree with Fodor and others that this is a perfectly coherent story of how, with respect to language uses, paradigmatic Frege-cases arise. But notice that for all I have said, Hammurabi's stargazing pet Sloughi (his Arabian Greyhound) may have found itself in more or less the same epistemic predicament, save the linguistic facts. For if one assumes (as I do) that any suitably sophisticated non-linguistic animal can think about the same object in different ways without thereby knowing/recognizing that they are in fact thinking about the same object, then Frege's puzzle is not a puzzle about language, *per se*, but rather a puzzle about the nature of *thought*.

To take a more realistic example, imagine Fido whose Master happens to be a postal delivery worker. It so happens that Master's mail route includes his own home. Thus, every morning Fido sees Master leave for work dressed in plain clothes just to return midday dressed in postal garb to deliver the day's mail. Suppose further that Master's wide-brimmed hat cast a shadow over his face such that from Fido's vantage point through the living room window he is unrecognizable as such, which is to say as Master. In these circumstances, it seems plausible enough to suppose that Fido is capable of thinking about the same object in two different ways without realizing it. Although, Fido might eventually intuit that Postman = Master; for instance if Master one day decided to pop in to say hello. The relevant point is that, in my view, not every instance of Frege's puzzle can be explained by appeal to facts about natural language. In turn, not every instance will involve a subject's metalinguistic-semantic competence (MSC), as argued below. By the same token, everyone agrees that with respect to Frege-subjects who also happen to be language users, a sincere and forceful utterance of a true natural language identity statement of the form 'α is β' *can* effect the same result—i.e., it seems

to be a brute fact that for reasons yet to be determined such utterances can cause a subject to revise his/her beliefs about how many objects/individuals are the contextually-salient bearers of the names ‘α’ and ‘β’. So qualified, given that Frege’s puzzle is typically cast as a linguistic phenomenon, which is how Frege himself conceived the puzzle, I will continue to restrict discussion to the causal efficacy of statements of the form ‘α is β’ as a linguistic means of disabusing Frege-subjects of their mistaken beliefs.

7.6. Addressing Frege’s puzzle

7.6.1. The general case

So qualified, Fodor’s way of solving Frege’s puzzle is equally straightforward. First, let us assume for the sake of argument that, in the general case, an utterer of ‘α is β’ is using the names ‘α’ and ‘β’ referentially, and the verb ‘be’ to express a substantive relation of numerical identity. Thus, let us grant with Frege from above that ‘α is β’ is understood as elliptical/shorthand for a more fully articulated sentence of the form ‘α is identical to β’, or ‘α is the same as β’, or some such equivalent statement. Whatever it is, I will also just assume for without argument that what the English verb ‘be’ (along with cross-linguistic cognates) *expresses* in this context of ‘α is β’ is something like the concept IDENTICAL-TO(x, y), or perhaps SAME-AS(x, y), which again by assumption expresses a substantive (i.e., meaningful) relation of numerical identity that holds between the referents of the names ‘α’ and ‘β’.

Granted these assumptions, we might represent the thought expressed by (10) as one of either (11) or (12):

(10) Hesperus is Phosphorous

(11) ∃x ∃y [(x=Hesperus) & (y=Phosphorous) & IDENTICAL-TO(x, y)]

(12) ∃x ∃y [(x=Hesperus) & (y=Phosphorous) & SAME-AS(x, y)]

However, the question I am most interested in is how a subject in Hammurabi's condition will *understand* an utterance of (10). As a baseline, let us assume for the moment that what Hammurabi represents in consequence of grasping the meaning of (10) is the thought/proposition in either (13) or (14):

(13) ∃x ∃y [(x=HESPERUS) & (y=PHOSPHOROUS) & IDENTICAL-TO(x, y)]

(14) ∃x ∃y [(x=HESPERUS) & (y=PHOSPHOROUS) & SAME-AS(x, y)]

For purposes of describing Hammurabi's psychological state, I will use the concept-names 'HESPERUS' and 'PHOSPHOROUS' to stand proxy for their representational contents such that: (x=HESPERUS) ≡ (x=the planet Venus *qua* Hesperus), and (x=PHOSPHOROUS) ≡ (x=the planet Venus *qua* Phosphorous). Notice again that while HESPERUS and PHOSPHOROUS are formally distinct mental representations, by assumption they are content-identical; they both "refer to" the object that is the planet Venus.

More specifically, granted assumptions about the semantics of proper names from above, if Hammurabi understands the name 'Hesperus' in an utterance of (10) as being used referentially, by hypothesis he will understand it as instruction to activate his concept_W IS-CALLED(HESPERUS, HESPERUS(w)). Likewise, if he understands 'Phosphorous' referentially, he will understand it as an instruction to activate IS-CALLED(PHOSPHOROUS, PHOSPHOROUS(w)). In turn, I am assuming that Hammurabi will understand the verb 'be' (or again its Akkadian equivalent) as expressing, let's say, the concept IDENTICAL-TO(x, y).

Now, what this suggests is that Hammurabi and others alike will understand an utterance of (10), *ceteris paribus*, as expressing the thought/proposition that Hesperus is [numerically identical to] Phosphorous. However, following Evans (1982) and others, it seems reasonable to assume that one cannot, in the general case, think about something as such without thinking about that thing in some way or another, which is to say under some description (or guise, or mode of presentation, or what have you). For example, thinking about Hesperus as the evening star is to think about Hesperus under a particular description. Thus, I will assume that activation of Hammurabi's concepts HESPERUS and PHOSPHOROUS in consequence of his interpretation of 'Hesperus' and 'Phosphorous' will normally result in the activation of beliefs such as (15) and (16) below:[25]

(15) $\exists x$ [x=HESPERUS & STAR(x) & EVENING-VISIBLE(x)]

(16) $\exists x$ [x=PHOSPHOROUS & STAR(x) & MORNING-VISIBLE(x)]

Keep in mind that at present we are still assuming that Hammurabi does not believe that 'Hesperus' and 'Phosphorous' corefer and, correlatively, that he is unaware of the fact that his concepts HESPERUS and PHOSPHOROUS are content-identical.

As a result, I take it that Hammurabi will initially be confused by his informant's utterance of (10). Specifically, given what Hammurabi believes to be mutually exclusive facts about Hesperus and Phosphorous, an utterance of (10) will strike him initially as being false if not incoherent. Alternatively, it might be the case that subjects in Hammruabi's position will understand (10) as expressing what Strawson (1950) referred to as *qualitative* identity between its two NP/DP-referents—i.e., that Hesperus and Phosphorous are merely qualitatively indistinguishable, which does not entail their

[25] I allow that activation of HESPERUS and PHOSPHOROUS might also normally be accompanied by activation of some sort of mental imagery of the planet Venus under different visual modes of presentation.

numerical identity. But in either case, as a way of bringing out the point more clearly consider (17) as addressed to Lois Lane:

(17) But Lois, Clark Kent *is* Superman

At least anecdotally speaking, one can imagine Lois pausing a moment, confusedly, and then replying: "Wait? Are we talking about the same Clark Kent?" This remark, in my estimation, is really just an abbreviated way of asking "Are we talking about the same person that I know by the name 'Clark Kent'?" In other words, in the face of an apparent absurdity my hypothesis is that Lois' thoughts will immediately turn metalinguistic. Yet this should be unsurprising given that her concepts$_W$ of the names 'Clark Kent' and 'Superman' are alit in virtue of having interpreted those names. Something similar can be said of Hammurabi.

Indeed, I suspect that Hammurabi's initial perplexity will cause him to briefly step back and consider the logical implications of the linguistic/grammatical form of his informant's utterance in search for clues as to how to make sense of it. This exercise can be revealing, however. First of all, given Hammurabi's belief that 'Hesperus' and 'Phosphorous' are type-distinct names he will not perceive his informant's utterance of (10) as an instance of Leibniz's law. And from this he can justifiably infer, given Gricean principles of charity, that it contains a potentially valuable piece of information.[26] Yet given that Hammurabi believes that 'Hesperus' and 'Phosphorous' do not corefer, (10) will appear to him to express a logical contradiction. As a competent speaker, however, Hammurabi also knows (at least tacitly) that nothing in the rules of his grammar (i.e., his I-language) prohibits their coreference. Given this much, Hammurabi can deduce from its

[26] Much of this should sound familiar from my discussion of Frege above, which is largely what justified my inclusion of that rather lengthy discussion in this chapter.

linguistic form that in order for the utterance to be true, it must be the case that the names ‘Hesperus’ and ‘Phosphorous’ are in fact being used in reference to the same object. And if true, this further implies that his concepts$_W$ HESPERUS(w) and PHOSPHOROUS(w) are concepts$_W$ of coreferential names. Now, should Hammurabi come to accept this much on authority of his informant’s testimony, he will be rationally compelled to conclude that there exists just one contextually-salient individual who is a (or perhaps the) bearer of both names, and not two as he previously believed.

In short, given his mutually exclusive beliefs about Hesperus and Phosphorous, Hammurabi has good reason to reject his informant’s assertion. At the same, Gricean principles push in the other direction—that he should assume that his informant is speaking truthfully and informatively. Indeed, if Hammurabi takes his informant to be knowledgeable and trustworthy, her testimony alone may be sufficient to justify his acceptance of her utterance as true. Yet whatever evidence Hammurabi brings to bear on his judgment, should he accept his informant’s testimony he will (or should) feel rationally obliged to bring his beliefs into alignment with hers regarding how many objects are referred to with the names ‘Hesperus’ and ‘Phosphorous’. In consequence, Hammurabi will also be compelled to accept that his concepts HESPERUS and PHOSPHOROUS are in fact content-identical. And in consequence of this epiphany, he will thereby be rationally compelled to integrate his preexisting beliefs about Hesperus and Phosphorous such that *whatever* he believes true (in a *de re* sense) about the one object must also be true of the other.

But even granted this much, I have not yet hazarded a guess as to precisely which property of (10), or perhaps rather the thought that it expresses, renders that statement

informative in the relevant sense. To repeat, Frege held that the cognitive value of an identity statement is determined by the thought that it expresses—a view designed to preserve the objectivity of thought. Thus, on Frege's view, to grasp the cognitive value of an identity just is to grasp the thought that it expresses. And so we might conclude that an utterance of (10) is informative to a subject in Hammurabi's epistemic condition only to the extent that he is capable of grasping the thought expressed. Counterfactually speaking, we can suppose that if Hammurabi were not in possession of the singular concepts HESPERUS and PHOSPHOROUS he would be incapable of thinking about either Hesperus or Phosphorous *as such*. And if he is incapable of thinking of thinking about Hesperus and Phosphorous as such, it would seem to follow that Hammurabi is thereby incapable of grasping the thought expressed by (10). And if he failed to grasp the thought expressed by (10), he would by Frege's lights be incapable of grasping its cognitive value. In turn, if Hammurabi failed to grasp either the thought expressed or its cognitive value, he would not *thereby* feel rationally obliged to revise his beliefs about how many objects are the bearers of the names 'Hesperus' and 'Phosphorus'.[27] If correct, it would appear that the informativeness of (10) *for Hammurabi* depends on his possession of HESPERUS and PHOSPHOROUS.

I wish to challenge this conclusion, however, by demonstrating that a competent speaker can perfectly well understand the *meaning* of an identity statement of the form 'α is β' without having singular concepts of the referents of the names 'α' and 'β'. If successful, I take this result to show that a competent subject can understand the *meaning* of 'α is β' without thereby grasping, or anyhow fully grasping, the thought that it

[27] Though he may have had independent reason for doing so.

expresses. I will at once attempt to demonstrate that an utterance of 'α is β' can be informative in the relevant sense merely by understanding its meaning. My hypothesis, in other words, is that subjects need not fully grasp the thought expressed by 'α is β' in order for its utterances to informative. And if that's correct, this result would at least *suggest* that the informativeness (or cognitive value) of a true identity statements of the form 'α is β' is *not*, as Frege held, determined (or at least not fully determined) by the thought that it expresses.[28]

As a way of motivating the view, I will first attempt to make the case with respect to the informativeness of simple predicative statements containing proper names. This result will then serve as a baseline for what I take to be general conditions on the informativeness of any declarative statement. I will then later attempt to run a similar argument specifically with respect to the informativeness of identity statements. But first I want to very briefly establish a methodological principle advanced by Strawson (1974) who argues that we should attempt to locate that which is informative for *any* competent speaker regardless of what they may or may not know/believe about the relevant objects of reference. For otherwise, the informativeness of an identity statement would, as Strawson (*ibid*: 44) observes, be listener-dependent:

> …different audiences, variously equipped, might learn various different things from one and the same identity statement. It might have different informative values, so to speak, for each of them.

He rejoins (*ibid*), however:

> But the question is really about the constant informative value, if any, which it has for everyone who learns anything from it.

[28] Alternatively, one might argue that while the cognitive value of the statement may in fact be *determined* by the thought expressed, one's *grasp* of the former does not depend on one's grasp of the latter.

In short, Strawson concludes that like that of ordinary predicative statements, the informativeness of identity statements cannot be listener-dependent. Thus, the informativeness of identity statements cannot depend on what listeners antecedently know/believe about the relevant object of reference.

Now, like Frege, Strawson seems to assume that the proposition expressed by an identity statement is that which determines its cognitive value, which is again a conclusion I wish to reject. However, as general methodological principle one can nevertheless agree with Strawson that what we want from a theory of informativeness is a theory of what *any* competent subject stands to learn from an identity statement insofar as they both understand what those statements mean and accept them as true. As an advertisement, my argument to this conclusion has the general shape of a *reductio*. As in the end, I will not be taking a firm stance on the question of what 'α is β' expresses other than to demonstrate that its utterance can be informative in the relevant sense for subjects who understand what it means yet fail to fully grasp the proposition expressed. Ultimately what we want is a causal explanation for why any competent speaker might, merely in virtue of having understood its meaning and having accepted his informant's statement as true, feel rationally persuaded to revise his belief about how many objects are under discussions. Cast in these terms, my conclusion is that the informativeness of an identity statement is tied to the logical and/or semantic implications of its linguistic form.

7.6.2. General conditions on informativeness

I would like to begin here by building a toy theory about how knowledge and/or concept acquisition occurs that runs parallel to remarks made in Chapter 3. But in this

case I focus on how we come to acquire new knowledge and/or concepts on the basis of expert testimony. For I assume that what such a theory will deliver is at once a theory of how, or in what way, utterances of true declarative sentences of natural language, including identity statements, can be *informative* for competent yet otherwise uninformed/misinformed interpreters. As a starting point, I take for granted that competent listeners can learn substantive (i.e., material) facts about the world simply by understanding and accepting as true the verbal testimony of knowledgeable and trustworthy informants—i.e., on the basis of what those who know better than us *say*.

A paradigm example of how we come to learn about the existence of things, for example, is by way of a demonstrative sentence such as (18):

(18) That is Hesperus

Suppose the addressee of (18) is Max, who by assumption has no clue as to who or what Hesperus is and has never before encountered the name 'Hesperus'. While at a conference of amateur astronomers, Max overhears the name 'Hesperus' mentioned in conversation. Out of curiosity he asks his friend Minnie, who Max takes to be a knowledgeable and trustworthy informant, "Who is Hesperus?" To which Minnie replies: "Well, first of all Hesperus is not a *who* but a *what*. But let me just show you." Minnie then marches Max onto the veranda, points to Venus shining brightly in the evening sky and utters (18).

Traditionally, philosophers of language would classify (18) as a "genuine" identity statement—i.e., a statement according to which the verb 'be' expresses a relation of numerical identity between its two NP-arguments (or DP-arguments, depending on one's

theoretical commitments).[29] However, many linguists these days classify (18) as an *identificational* copular sentence, or ICS for short. Eliding several important details, ICSs are so-called because they provide addressees with a linguistic means of identifying (i.e., picking out, referring to, etc.) the object referred to by the grammatical subject. Specifically, according to this conception Minnie is using the demonstrative 'that' in (18) *referentially* and the name 'Hesperus' *predicatively*. And what she thereby expresses with her utterance is the proposition that the object demonstrated *is called 'Hesperus'*. That is, Minnie is using (18) to at once refer to the planet Venus and to share with Max a customary way of identifying that object by name.

Now, however we classify it, if any statements of natural language are informative they surely include Minnie's utterance of (18). For by accepting her utterance as true Max will have learned about the existence of a hitherto unknown object. In consequence of his experience, Max will have also acquired a crude mental model of Hesperus (e.g., that it is the brightest object in the evening sky) that will aid in his describing and/or re-identifying Hesperus in the future. In fact, it strikes me as plausible to suppose that this experience alone is sufficient for Max to acquire the singular concept HESPERUS, which will henceforth allow him to think about Hesperus *as such*, which is to say the planet Venus *qua Hesperus*. In short, it seems quite clear in this case that Max has acquired a substantive piece of knowledge about the world, and hence that Minnie's utterance of (1) is informative in the relevant sense.

---

[29] Many language theorists nowadays accept some version of Abney's (1987) DP-Hypothesis according to which all NP projections are dominated by a determiner, D. These NPs therefore occur as complements of a higher determiner phrase (DP) projection. With respect to proper names, I am again inclined to follow Burge (1973) according to which NPs headed by bare proper names (in English) are dominated by a *covert* D which behaves semantically like a demonstrative element more or less equivalent to the English 'that'. Specifically, we might think of this covert D as playing the role of a referential index in an LE such as [$that_1$ [$_{PN}$ Aristotle]].

In addition, I take it Max's experience is sufficient to have acquired the name 'Hesperus'. I further assume, as argued back in Chapters 3 and 4, that Max will have at once acquired a concept$_W$ of that name, which is to say HESPERUS(w). He will then have at once used HESPERUS(w) to form the explicit metalinguistic belief in (19), which can be read as stating that there is Word (or name), w, that is 'Hesperus':

(19) ∃w [HESPERUS(w)]

Moreover, given his core linguistic competence I take it that in addition to having become directly acquainted with its referent, Max will in a sense know *a priori* what the name 'Hesperus' means. Specifically, he will know that 'Hesperus' expresses the property of being called 'Hesperus'. Or put in conceptual terms, Max knows that 'Hesperus' expresses the metalinguistic concept$_W$ IS-CALLED(HESPERUS(w)). Of course, since what Max learns in virtue of having acquired the name 'Hesperus' is purely metalinguistic in nature, this would not, for reasons discussed in Section 7.3, pass muster by Frege's criterion of cognitive significance. And with this much I can agree. But I again take as non-tendentious that an utterance of (18) is informative in Frege's sense of making contact with the relevant object of reference.

In addition to demonstrative constructions such as (18), other simple predicative statements such as (20) are also straightforwardly informative in the relevant sense:

(20) Hesperus is a planet

For suppose that Max, who is now duly acquainted with Hesperus and thus duly equipped with the concept HESPERUS, and now knows Hesperus by name, wants to know what sort of celestial object Hesperus is. Max once again turns to Minnie for answers and who dutifully responds with an utterance of (20). Let us also assume that Max knows what

planets are, and therefore has possession of the concept PLANET(x).[30] So equipped, Max will presumably understand Minnie's utterance of (20) to have expressed the thought that Hesperus is a planet. Under the further assumption that he accepts Minnie's testimony as true, Max will have thereby acquired a true belief whose content can be represented as follows:

(21) ∃x [x=HESPERUS & PLANET(x)]

Here again for purposes of describing a believer's psychological state, I am using the concept-name 'HESPERUS' to stand proxy for its content. Also here again I take as non-controversial that Max's acceptance of Minnie's utterance constitutes a significant extension in his knowledge of the world.

To generate a more tendentious example, we need only reverse the order of events above such that Max first encounters an utterance of (20) followed by (18). That is, as before Max overhears the name 'Hesperus' mentioned in conversation and asks "Who is Hesperus?" However, this time rather than immediately demonstrating Hesperus, Minnie instead attempts to first *describe* Hesperus to Max by asserting (20). The question is again whether in this context Minnie's testimony is informative, and if so in what way? In this case, I take it that what Max will have learned from Minnie's utterance is that something called 'Hesperus' is a planet, which I will regiment as follows:[31]

(22) ∃w ∃x [HESPERUS(w) & IS-CALLED(x, HESPERUS(w)) & PLANET(x)]

In rough paraphrase, (22) reflects Max's belief that there is Word, w, which is 'Hesperus', and for some x, x is called 'Hesperus' and x is a planet. Observe, however,

[30] Though I think my main point here could be made more dramatically by assuming that Max arrived at the table as a blank slate. Yet for brevity, the present example should do the trick.

[31] As deployed in (5), IS-CALLED(HESPERUS(w)) is relational variant of the monadic concept$_w$ IS-CALLED(HESPERUS(w)), both of which I assume competent speakers possess.

that while Max has clearly learned something, given that (22) is an essentially metalinguistic belief it would again presumably not count as informative by Fregean standards. In particular, I suspect that neo-Fregeans would argue that since by assumption Max lacks the singular concept HESPERUS he is incapable of forming and thus grasping the singular thought *expressed* by (20). And if Max fails to fully grasp the thought expressed by (20) he cannot fully grasp its cognitive value.

I plan to challenge this last claim presently. But to aid in that effort suppose next that Max wants to know precisely *which* planet is called 'Hesperus'. Again turning to Minnie for help, she marches Max onto the veranda and confidently asserts (18) (i.e., "Look up there, Max. You see? *That* is Hesperus"). Here again I take there to be no disagreement that Minnie's utterance of (18) is informative. For under present assumptions Max will become directly acquainted with the planet Venus *qua Hesperus* and thereby acquire the concept HESPERUS. In this case, however, Max antecedently knew what the demonstrated object is called. He also knew that whatever is called 'Hesperus' is a planet. Thus, in consequence of Minnie's utterance of (18) Max will have learned, in a sense without being told, that Hesperus is a planet.

Now, we have already decided that for Max to learn that Hesperus is a planet constitutes a substantive extension of his knowledge. Furthermore, since we have tentatively decided that Minnie's utterance of (20) is *un*informative, one assumes that the significance of Max's coming to learn that Hesperus is a planet must owe solely to Minnie's utterance of (18). Yet notice that while Max's acceptance of (18) contributed to his acquisition of HESPERUS, and thus doubtlessly contributed to his learning that Hesperus is a planet, (18) neither expresses nor implies the fact that Hesperus is a planet.

Thus, (18) alone cannot explain Max's having learned that Hesperus is a planet. Rather, it seems that *both* (18) *and* (20) contributed to Max's learning that Hesperus is a planet. But if that's right then (20) must be informative after all!

This conclusion strikes me as correct, and here is my conjecture as to why. By accepting (20) as true on the basis of Minnie's authority, Max has undertaken a new ontological commitment regarding what there is, which is to say *how many* things exist in the world (which is one more thing than he believed prior to Minnie's statement). Hence, there is an intuitively clear sense in which what Max has learned in virtue of his acceptance of (20) is in fact "object-involving" and therefore satisfies the spirit of Frege's criterion of cognitive significance. Moreover, Max initially acquired this first piece of knowledge without benefit of a stable concept of Hesperus. As such, prior to his acquisition of HESPERUS Max was evidently unable to fully grasp the thought expressed by (20). Nevertheless, with respect to both (18) and (20) Max was capable of understanding the *meaning* of Minnie's utterances. And so what this suggests, in short, is that Frege's criterion of cognitive significance is too strong, or at least with respect to simple predicative statements. Yet as I argue below the same sort of argument can be run with respect to identity statements to the same conclusion: that Frege-subjects need not fully grasp the thought expressed by true identity statements for those statements to be informative *insofar as subjects understand what they mean*.

7.6.3. Strawson on informativeness

As mentioned above, I am adopting from Strawson (1974) the general methodological principle that what we want from a theory of the informativeness of identity statements is a theory of what *any* competent listener stands to learn from those

statements insofar as listeners both understand what those statements mean and accepts them as true. In brief, while Strawson eschews talk of concepts, his proposed solution to Frege's puzzle tracks closely with that of Fodor's as presented above. Specifically, Strawson makes reference to the notion of mental files around which we form "clusters or bundles" of "uniquely identifying knowledge" about the relevant objects of reference. However, Strawson (*ibid*: 33-4) observes:

> But of course there is no reason in the world why a hearer should not be in possession of, as it were, segregated bundles or clusters of identifying knowledge which are in fact, unknown to him, bundles or clusters of identifying knowledge about the same thing. If this is so, and if one name in an identity statement invokes one such cluster, while the other invokes another, then indeed the hearer will learn something, and not simply something about expressions, from the identity-statement which couples the two expressions.

Strawson's proposal about the informativeness of identity statements is tied up with his notion of having "command" of a proper name. Specifically, and in contrast to my account from Section 7.6.2, he maintains that in order to understand the meaning of a declarative sentence containing referential uses of proper names, and hence to grasp the thoughts they express, interpreters must have command of the names in question. And by Strawson's lights it is sufficient to have command of a name if that name evokes "some kind of identifying knowledge, in the hearer's possession, of the bearer of the name." Unfortunately, Strawson fails to specify exactly what counts as "some kind of identifying knowledge," as this could mean several things. Yet notice that he also does not specifically say that such knowledge must be *uniquely* identifying. Indeed, given what Strawson says elsewhere he seems to allow for the possibility that subjects may only be acquainted with the object of reference by *indefinite* description, which is a rather weak condition (more on this below).

Given my account of the semantics of proper names, however, even this condition is too strong. As again on my view competent interpreters can understand what a name means, linguistically speaking, without having any kind of discriminating knowledge about its customary referent. Nevertheless, I can accept Strawson's rather weak constraint on having command of a name to make the same point. And Strawson's point is again that the informativeness of 'α is β' is constant, not variable. That is, identity statements do not depend for their informativeness on differences between what particular subjects may or may not know/believe about the relevant object(s) of reference. Rather again, as Strawson urges, what we want is theory of what *any* competent subject stands to learn from their understanding and acceptance of 'α is β'. My aim throughout this chapter has been to draw out, bit-by-bit, various features of identity statements that might help us determine what I will call the *lowest common denominator* (LCD)—that property (or set of properties) of identity statements that render them informative for anyone who understands their meanings and accepts them as true.

Ultimately, however, my conclusion differs from Strawson's. As again it seems to me that a subject need not grasp the thought/proposition expressed by an identity for that statement to be informative in the relevant sense. Rather, on my view the LCD principle has to do both with what subjects know about the meanings of proper names together with what they can justifiably infer from statements that contain those names on the basis of their logical/linguistic forms. And as mentioned in the beginning, I take it that what any competent speaker can justifiably infer from 'α is β' is that there exists at most one contextually-salient bearer of both 'α' and 'β'. I motioned in the direction of how a duly acquainted subject such Hammurabi might land on this thought. However, as a way of

motivating Strawson's dictum I next examine a degenerate case where by assumption Frege-subjects know little more about the relevant objects of reference than their names. And this is to suppose that our subjects lack concepts of those objects, and are therefore incapable of representing/grasping the thought expressed by statements that contain referential uses of their names. Yet as argued in Section 7.6.2, these subjects are, in my view, nevertheless capable of understanding what those names mean, linguistically speaking.

7.6.4. A degenerate case

First consider Kripke's famous Feynman/Gell-Mann example according to which a competent subject, again call him Max, only vaguely recalls having heard of two people who go by the names 'Feynman' and 'Gell-Mann'. Apart from their names, however, Max only remembers hearing that both individuals are famous physicists. That is, by assumption Max can neither uniquely describe nor visually identify either individual. He obviously cannot, therefore, discriminate one individual from the other except by name. Under these circumstances, I think it's safe to assume that Max lacks singular concepts of Feynman and Gell-Mann, which is to say FEYNMAN and GELL-MANN. As before, then, it seems safe to assume that Max is therefore incapable of thinking about either Feynman or Gell-Mann as such. However, according to Kripke Max can nonetheless succeed in using 'Feynman' and 'Gell-Mann', respectively, to refer to Feynman and Gell-Mann by name insofar as he intends to use those names with their customary reference. Moreover, if I understand Strawson correctly we can say that Max has "command" of those names, though again in the weak sense of merely possessing *some kind of identifying knowledge*; *viz.*, that both are famous physicists.

More to my point, given that Max knows that the names ‘Feynman’ and ‘Gell-Mann’ are names of famous physicists, it is not implausible to suppose that he has both recorded those names in his mental lexicon and also acquired concepts$_W$ of those names; call the latter FEYNMAN(w) and GELL-MANN(w). In my view, it would also be unsurprising if, for lack of an alternative, Max has used his concepts$_W$ FEYNMAN(w) and GELL-MANN(w), respectively, to form the metalinguistic beliefs that both ‘Feynman’ and ‘Gell-Mann’ are names of famous physicists, which I will represent as (23) and (24):

(23) ∃w ∃x [FEYNMAN(w) & IS-CALLED(x, FEYNMAN(w)) & PHYSICIST(x)]

(24) ∃w ∃x [GELL-MANN(w) & IS-CALLED(x, GELL-MANN(w)) & PHYSICIST(x)]

Here again, while Max may be incapable of uniquely identifying Feynman and Gell-Mann, nor distinguishing the one man from the other, as a competent speaker he knows what the names ‘Feynman’ and ‘Gell-Mann’ *mean*, as expressed by his concepts$_W$ IS-CALLED(FEYNMAN(w)) and IS-CALLED(GELL-MANN(w)), respectively. In terms of comprehension, Max will therefore understand the meanings of the names ‘Feynman’ and ‘Gell-Mann’, respectively, as instructions to activate his concepts$_W$ IS-CALLED(FEYNMAN(w)) and IS-CALLED(GELL-MANN(w)). In turn, activation of FEYNMAN(w) and GELL-MANN(w) will *ceteris paribus* at least prime the activation of Max’s metalinguistic beliefs represented by (23) and (24) (i.e., that ‘Feynman’ and ‘Gell-Mann’ are names of famous physicists). In short, this sequence of interpretive events would explain why Max’s interpretation of the names ‘Feynman’ and ‘Gell-Mann’ lead to thoughts about famous physicists despite lacking the singular concepts FEYNMAN and GELL-MANN.

Now, to put a hypothetical twist on Kripke's example, suppose it turned out that, unbeknownst to Max, that there is just one famous physicist—Feyn-Mann—who is called both 'Feynman' and 'Gell-Mann'. Suppose further that at a physics convention Max overhears the names 'Feynman' and 'Gell-Mann' mentioned in conversation. Wishing to participate in the discussion, Max naively interjects: "So, who you do think has done more to advance modern physics, Feynman or Gell-Mann?" To which one of the conversational participants replies:

(25) What do you mean, Feynman is Gell-Mann!

As pursued above, the relevant question here is whether Max learns anything of substance merely by understanding [the meaning of] his respondent's utterance of (25)? My claim is that he does.

First of all, let us again assume for the sake of argument that Max understands the names 'Feynman' and 'Gell-Mann' in (25) as having been used referentially, and as charitably that he understands the verb 'be' in this context as expressing a concept of numerical identity. Given that Max is incapable of thinking about Feynman/Gell-Mann as such, by assumption he is unable to fully form the thought *expressed* by (25). Rather, under present assumptions what he will understand from his respondent's utterance is that 'Feynman' and 'Gell-Mann' are names of the same person—*Feyn-Mann*. And from this thought he can justifiably infer that in order for (25) to be true it must be the case that there is just one contextually-salient individual under discussion, though again Max does not know precisely who that individual is.

In purely psychological terms, my hypothesis is that Max's interpretation of (25) will cause activation of his concepts$_W$ IS-CALLED(FEYNMAN(w)) and IS-CALLED(GELL-

MANN(w)), which again *ipso facto* involves activation of his concepts$_W$ of the names 'Feynman' and 'Gell-Mann', which is to say FEYNMAN(w) and GELL-MANN(w). That is, by merely understanding the meaning of (25), Max's thoughts will naturally gravitate toward the metalinguistic aspects of the utterance, as by assumption he lacks the singular concepts FEYNMAN and GELL-MANN. However, as with Hammurabi from above Max can nonetheless read quite a lot off of the linguistic form of the utterance. Glossing past certain details, Max will know more or less *a priori* that in order for his interlocutor's utterance to be true, it must be the case that the names 'Feynman' and 'Gell-Mann' corefer. This further implies that his concepts$_W$ FEYNMAN(w) and GELL-MANN(w) are, contrary to prior beliefs, concepts$_W$ of coreferential names. Moreover, should Max come to accept this much on authority of his informant's testimony, he will be rationally compelled to conclude that there exists just one contextually-salient individual who is a (or perhaps the) bearer of both names, and not two as he previously believed.

Here again the upshot is that merely by understanding the meaning of (25), and accepting it as true, Max will feel rationally compelled to revise his ontological commitments regarding how many individuals are the bearers of 'Feynman' and 'Gell-Mann'. Now, if you bought the argument of Section 7.6.2 above, then I think you ought to agree that what Max has learned from his acceptance of (25) constitutes a substantive advance in his knowledge of the world. And if that's right, then identity statements can be informative for competent subjects despite not fully grasping the thoughts/propositions that those statements express.

As a way of cementing the point, suppose that Max becomes intrigued by this famous physicist and goes on to learn more about him. Eventually, Max acquires a

singular concept of this individual; call it FEYN-MANN (recall that for illustrational purposes we are imagining that there exists just one person called by the names 'Feynman' and 'Gell-Mann'). Since Max antecedently believes that *someone* called 'Feynman' and *someone* called 'Gell-Mann' are famous physicists, in virtue of having acquired FEYN-MANN he will in a sense "automatically" know that the referent of FEYN-MANN, which is to say Feyn-Mann, is a famous physicist. In other words, in consequence of his acquisition of FEYN-MANN, together with his acceptance of (25) above, Max will have formed the following *singular* belief:

(26) $\exists x\,[x=\text{FEYN-MANN} \,\&\, \text{PHYSICIST}(x)]$

I again take as non-controversial that the thought/belief represented by (26)—i.e., that Feyn-Mann is a physicist—constitutes a substantive piece of knowledge that Max did not possess prior to his acceptance of (25). Rather, it is only taken together that Max's beliefs in (23) and (24) that his interlocutor's utterance of (25) culminated in Max's belief in (26). It therefore seems clear that his interlocutor's utterance of (25) was in fact informative in the relevant sense.

Notice, however, that an utterance of (25) is informative only to the extent that Max recognizes the metalinguistic-semantic information that it pragmatically (or linguistically) conveys. In other words, my claim is that the informativeness of (25) rests *not* with Max's ability to *identify who* or *what* is being referred to with 'Feynman' and 'Gell-Mann'. Rather, the informativeness of the statement lies in Max's recognition of *how many* objects are being referred to with those names. But to grasp this thought required first recognizing that in order for the statement to be true it must be the case that 'Feynman' and 'Gell-Mann' are being used in reference to the same individual.

Notice further that it cannot be the case that the thought expressed by (25) (i.e., that Feynman is Gell-Mann) is logically equivalent to the thought there is just one contextually-salient bearer of the names 'Feynman' and 'Gell-Mann'. For as just demonstrated, a subject such as Max who lacks the concepts FEYNMAN and GELL-MAN appears capable of grasping the latter thought but not the former. The latter thought in turn, if accepted as true, appears sufficient to rationally compel subjects such as Max to revise their beliefs about how many objects are under discussion. In Hammurabi's case from earlier, given that he harbors mutually exclusive beliefs about Hesperus and Phosphorous he may be more reluctant to accept an utterance of (10) above as true (i.e., "Hesperus is Phosphorous") than Max is to accept (25) on testimony alone. But generally speaking, what one *understands* by the utterance and one's *justification* for accepting it as true are independent considerations. As ultimately I think that Max and Hammurabi, despite being in very different epistemic positions, derive basically the same conclusion, which is that there is just one object that is a/the bearer of two names. If correct, this suggests that the thought speaker express with an identity statement may not match the thought understood by competent yet otherwise misinformed addressees.

Of course, a speaker would not normally assert an identity statement unless he/she believes that there is confusion among addressees about how many objects/individuals are under discussion. It might therefore be said part of what speakers are trying to communicate is precisely what any competent listener stands to learn from their utterances, which to repeat is that there is just one object/individual under discussion. Moreover, it might be said that identity statements of the form 'α is β' are not only informative but also carry *normative force* along the lines of a recommendation or, more

strongly, a *command*. For a speaker of, say, (10) is at once urging his listener to bring her beliefs into alignment with his own regarding how many objects are the respective bearers of the names in question. Thus, one can agree with Frege that (10) expresses a substantive thought about the relevant object of reference. However, *pace* Frege, my view is that the ability to *recognize* the cognitive significance of such statements requires the deployment of what in earlier chapters I called a speaker's metalinguistic-semantic competence (MSC) regarding his/her understanding/beliefs about what the names in question mean.[32]

In general, my claim is that the informativeness (or "cognitive significance") of natural language identity statements of the form 'α is β' is governed by the following principle:

> **(INF)ormativeness:** A true identity statement of the form 'α is β', where 'α' and 'β' are lexical expressions that contain coreferential uses of proper names, is *informative* just in case its utterance expresses or otherwise pragmatically conveys information that is sufficient to rationally compel a linguistically competent yet otherwise misinformed subject who understands its meaning, and believes it true, to revise his/her mistaken beliefs about *how many* objects are the bearers of those names. Parallel remarks apply to statements of the form 'α is *not* β'.[33]

On my view, the relevant information conveyed by an utterance of 'α is β' is the *metalinguistic fact* that the names 'α' and 'β' corefer. A subject's grasp of this proposition will in turn license what I above called the *lowest common denominator* (LCD)—that which any competent speaker stands to learn from a statement simply by

[32] The sort of MSC alluded to here is effectively what Fiengo & May (2006) refer to as "beliefs of Assignment." The main difference is that I have postulated a conceptual mechanism in attempt to explain how such beliefs are internally represented, processed, and the structural relation between the linguistic meanings of Words and corresponding metalinguistic concepts$_W$ of those Words.

[33] As mentioned in a similar footnote in Chapter 1, (INF) is forwarded as an epistemological or pragmatic principle, not a semantic principle.

understanding what it means and accepting it as true. Specifically, I take the LCD to be the *substantive* (i.e., *material*) inference that there exists just one (or at most one) contextually-salient object/individual that is the bearer of both names (or two in the case of 'α is not β'). While Frege-subjects who are suitably equipped with relevant concepts/beliefs such as Hammurabi may stand to learn something more from their acceptance of true identity statement as true, my strategy, following Strawson, has again been to seek the LCD—that which is common to all subjects who understand what those statements mean.

To be clear, I am not claiming that this is the *only* way of bringing such facts to light. Indeed, rational agents may well arrive at the same conclusion, as it were, by simply connecting the dots; i.e., by adducing relevant *extralinguistic* evidence in the same way that one might, as Frege thought, independently discover "that a new sun does not rise every morning, but always the same one." In general, I am again prepared to believe that even certain non-human animals can be Frege-subjects. But it's important to keep in mind that I am forwarding an empirical hypothesis of how competent language users utilize inferential relations licensed by the linguistic forms of certain natural language sentences in attempt to grasp what their utters are trying to communicate. Importantly, I take myself as having arrived at this conclusion without supposing, as do Fiengo & May, Strawson, along with most other commentators, that the relevant information specified by (INF) is part of the *thought expressed*.

Before concluding this chapter I wish to just briefly demonstrate how the proposal on offer generalizes to Kripke's Paderewski-cases.

7.6.5. Paderewski revisited (just briefly)

Recall from Chapter 6 that the special problem created by Paderewski-cases occurs when the names flanking the copula *appear* to be type-identical, as in (27):

(27) Paderewski is Paderewski

In this case, an interpreter's judgment about the proposition expressed more crucially depends on his judgment as to whether the lexical expressions (LEs) involved are type-identical or type-distinct, and correlatively whether they necessarily corefer (or not), as determined in part by whether those LEs are coindexed (or not). As in the previous chapter, let us assume here that Peter believes that there are two numerically distinct individuals, both who go by *the* name 'Paderewski'. Finding full agreement with Fiengo & May on this point, it seems to me plausible to suppose that given Peter's doxastic predilections he will naturally interpret (27) as involving type-distinct (non-coindexed) LEs which he believes do not corefer. Indeed, for Peter to interpret an utterance of (27) otherwise would be to charge his interlocutor with having uttered a triviality. Thus in order to make sense of its utterance, it must first occur to Peter (if only tacitly) that (27) might have the following (partial) grammatical form:

(27') $[_{NP}\ [_{PN}$ Paderewski$][x_1]]$ is $[_{NP}\ [_{PN}$ Paderewski$][x_2]]$

And from this revelation I assume that Peter's interpretation of (27/27') will proceed exactly as Hammurabi's interpretation of (10) from above. For like Hammurabi, Peter tacitly knows that nothing in the grammar precludes the coreference of type-distinct LEs. And if he finds independent reason to trust the testimony of his informant, he will thereby be forced to conclude that there is just one contextually-salient bearer of the non-coindexed LEs '$[_{NP}\ [_{PN}$ Paderewski$][x_1]]$' and '$[_{NP}\ [_{PN}$ Paderewski$][x_2]]$'.

I have dashed through this example pretty quickly. Yet given the extension discussion from the last chapter, and that of the present, I trust that the key claim is clear enough.

7.7. Chapter summary

To repeat a claim mentioned in the beginning, on my view Frege's puzzle about the informativeness of identity statements involving type-distinct coreferential names, and correlatively failures of substitution of those names *salva veritate* in opaque contexts, both arise for one of two simple reasons. Either (i) subjects do not know/believe *de dicto* that the names in question corefer (or not), or (ii) they have confused *de re* one individual for two (or *vice versa*). These two facts are intimately related, however, in the sense that ignorance of the one *explains* ignorance of the other.[34] For in the general case, to believe of two names that they do not corefer is *ipso facto* to be committed to the belief that those names have numerically distinct bearers.[35] Conversely, to believe *de re* that there exists two objects/individuals that are the respective bearers of type-distinct names rationally entails the subject's *de dicto* belief that those names do not corefer.

Similarly, there at least two ways out of the confusion. Subjects might either (i) learn, perhaps by being told, that the names in question corefer (or not), or (ii) infer from independent evidence that there exists at most one object that happens to be a (or perhaps *the*) bearer of both names (or in the reverse direction that there exists multiple bearers of the same name). These latter two facts are also related in the formal sense of providing

[34] To be clear at the onset, I assume that suitably sophisticated non-linguistic animals can in principle be Frege-subjects. If correct, then Frege's puzzle is not a puzzle about language, *per se*, but rather a puzzle about the nature of *thought*. Tis point will become clearer below. However, given that the puzzle is typically cast as a linguistic phenomenon, and which is how Frege himself conceived the puzzle, I will evaluate the phenomenon with respect to language users.

[35] By "general case" I mean to set aside instances of so-called "empty" proper names.

logical premises that connect one to the other such that a competent speaker who both recognizes and accepts the one fact will (or should), again *ceteris paribus*, feel rationally compelled to accept the other.

Yet as has been the focus of attention in the chapter, there is by common assumption a third way of becoming so-informed, which is to be told by way of a true identity statement of the form 'α is β' that α and β are in fact one and the same object/individual. And so far as I can tell, most contemporary researchers agree that with respect to natural language statements of this form constitutes the core *explanadum* of Frege's puzzle. So constrained, I have attempted to uncover which properties of such statements render them informative in the relevant sense. According to Frege, the "cognitive value" is determined by the thought/proposition that it expresses.

However, I have argued here that competent yet otherwise uninformed/misinformed language users need not fully grasp the thought expressed to grasp the informativeness of a true identity statement of the form 'α is β'. And while I have also tried to hedge my bets somewhat, if pressed for a firm answer as to what makes property of an identity statement renders it informative for any subject who understand what it means, my temptation is to say the metalinguistic fact that the names 'α' and 'β' corefer. As a subject's grasp of this proposition licenses what I take to be the central informative feature of identity statements, which is the fact that there exists at most one contextually-salient object/individual that is the bearer of both names (or two in the case 'α is not β'). As all of this fits in with the core thesis of this dissertation, the kind of competence I have been ascribing to Frege-subjects who grasp the informativeness of identity is chiefly what I called in earlier chapters their metalinguistic-semantic competence (MSC).

In short, I strongly suspect that anyone who has ever given Frege's puzzle more than a moment's thought has been seduced by Frege's intuitions in the *Begriffsschrift* that the informativeness of '*a* = *b*' is somehow grounded in the metalinguistic information that its utterance conveys. I have here attempted to demonstrate that Frege's *Begriffsschrift* intuitions were more or less sound, or again at any rate with respect to natural languages along with the help of a more empirically plausible theory of the linguistic meanings of proper names (i.e., one that makes no direct appeal to Fregean senses). I have purposely avoided related questions that arise with respect to so-called "Mates cases" (Mates [1950]) and what Jennifer Saul calls "simple sentences," mainly because I have not had the time to more carefully consider what role metalinguistic-semantic competence might play in the interpretation of such expressions. For this reason, I have elected to set these cases aside for further study. By contrast, for reasons of time and space, readers will have noticed that I have also swept aside the phenomenon of substitution failure of coreferential names in the context of belief reports. However, I trust that my target audience will have little trouble envisioning my response to this question as well. As in my view substitution failure in opaque contexts and the informativeness of identity statements are quite obviously related phenomena, though I won't pursue that argument here.

In the final chapter of this dissertation—Chapter 8—I will summarize the central claims of this study while by reiterating the explanatory role of MSC in utterance and/or expression interpretation. I will also briefly identify related avenues of future research.

8. Summary and Conclusions

8.1. Introduction

I positioned this study back in Chapter 1 as a theory of semantic competence, which broadly construed I have assumed to be a theory of whatever it is that one must know, or cognize, or otherwise mentally represent in order to be a semantically competent speaker of a natural language. More specifically, against the backdrop of a Chomskyan conception of natural language I began with a characterization of semantic competence as a speaker's ability to *understand* the expressions of his/her native *I-language*. As postulated by Chomsky, an I-language is an effective computational procedure implemented in the human mind capable of generating an endless array of meaningful linguistic expressions from a finite stock of lexical constituents according to a recursively specified set of grammatical rules. Expressions thus-generated must in turn satisfy legibility conditions (conditions on interpretation) at the interfaces to language-external consumer systems. When these conditions are met, understanding occurs (for the most part, but as qualified below).

Most relevant to the present topic, generable expressions must be interpretable at the interface to conceptual-intentional systems (C-I). By hypothesis, interpretable expressions bear certain semantic properties—which following Chomsky is to say *meanings*—that constrain their possible interpretations at the interface to C-I. More specifically, I have argued that linguistic meanings guide and constrain without fully determining which extralinguistic *concepts* are coherently expressible with which expressions across the various contexts in which those expressions are customarily deployed. The net effect, from a user perspective, is that linguistic meanings impose

constraints on what their host expressions can (and cannot) be used to describe, denote, or otherwise talk about in ordinary discourse. With respect to comprehension, I have also argued alongside Chomsky that we should think of linguistic meanings as *instructions* to activate semantically related concepts. At the lexical level, the meanings of individual Words are normally understood as *instructions* to C-I to activate one of possibly several basic/atomic concepts. Syntactically complex expressions, on the other hand, typically correspond to the activation of complex (i.e., internally structured) concepts, and at the sentential level to complete thoughts/propositions. Or as Paul Pietroski conceives them, the meanings of syntactically complex expressions are understood as, and indeed *just are* instructions to build/assemble complex monadic concepts by *fetching and conjoining* formally atomic monadic concepts (see Chapter 5 for discussion).[1]

Characterized in roughly this way, linguistic understanding is nothing more (and nothing less) than the ability to extract relevant linguistic properties from incoming speech signals in order to generate meaningful expressions—i.e., expressions that are *semantically interpretable* by language-external consumers. By "semantically interpretable," I merely mean that expressions *qua* meaning-instructions are in principle *executable* by their consumers. Notice, however, that instructions can often be executed, which is to say *satisfied*, in many ways. To borrow an illustrative analogy from Pietroski (p.c.), suppose I instruct you to fetch a box from a room. It so happens that inside this room you find several boxes of various proportions and colors. Of course, your awareness of extralinguistic context may influence your judgment as to what sort of box

[1] I have for the most part used the term 'activate' throughout as opposed to 'fetch' in order to remain relatively neutral about the exact nature of semantic compositionality. That said, I am attracted by the elegance and explanatory power/reach of Pietroski's rich and empirically supported elaboration of Chomsky's basic notion.

would best suit my needs in the situation at hand. For instance, if you know that I am in a rush and that any size/color box will do, you might just grab the first box in sight. And given that my instruction was in this respect non-specific, then strictly so long as you return *a box* you will have satisfied (i.e., executed) my request as instructed. By analogy, if the meaning of the Word 'book' is an instruction to fetch a concept of books (or book*ings*), so long as there is at least one such concept at the specified address then the meaning-instruction encoded by the lexical entry for 'book' is in principle executable and hence understandable (although performance limitations may, on occasion, disrupt its proper execution).

Notice further that to say that an expression *qua* meaning-instruction is executable is no guarantee that its execution will result in the activation/construction of a *comprehensible* thought. For instance, while the sentence "Colorless green ideas sleep furiously" conjures an unintelligible thought, on the present view it conforms to principles of grammatical well-formedness and is therefore meaningful (understandable/executable) in the relevant sense. That is, from a purely formal standpoint this expression is "legible" at the interface to C-I. And again from a strictly Chomskyan perspective, this is all that's required to understand expressions thus-generated and to thereby be a competent language user. In short, on this view it is the computational resources of a speaker's I-language that underwrites his/her linguistic competence, and specifically his/her *linguistic-semantic* competence (LSC). Thus, to possess an I-language is to possess a certain degree of LSC, where the "linguistic" modifier is added to remind ourselves that the competence in question is underwritten by a domain-specific faculty of the human mind dedicated to language processing. And to

qualify LSC as "semantic" is merely a reminder of which aspect of linguistic competence we are talking about.[2]

LSC is to be contrasted with what in Chapter 1 I labeled a speaker's *conceptual-semantic competence* (CSC), on the one hand, and *pragmatic-semantic competence* (PSC) on the other. There I characterized CSC as comprising our general conceptual knowledge of things in the world that we typically use language to communicate about. For example, knowing that dogs are quadrupeds (or anyhow typical ones) constitutes part of one's CSC, whereas understanding what the word 'dog' *means* is part of one's LSC. In contrast, to recognize that by asserting the sentence "The dog is loose again!," father is urging me to return Fido to his cage is part of my PSC. Explaining how general conceptual and pragmatic factors influence speaker-judgments about, say, the thought expressed by linguistic expressions is, on my view, an interesting question for philosophers of mind, language, and cognitive science. However, explaining the nature and structure of linguistic meaning is largely the purview of theoretical linguistics, and formal semantics in particular (although there will always be a degree of cross-disciplinary overlap in theoretical interests/purposes).

As a philosopher of language, my present interest lies primarily with how empirical findings might inform traditional philosophical questions about the relationship between language and thought, meaning and truth, and with respect to the present study the relationship between semantic theory and semantic competence. And again in my view a theory of core semantic competence, and LSC in particular, should be guided by our best

[2] Keep in mind that *ex hypothesi* the phonological properties of expressions are also *interpreted* (as instructions) by articulatory-perceptual (or sensorimotor) systems. Thus, phonological *understanding*, and hence phonological *competence*, plays an important role in a general theory of linguistic competence as well (not to mention what one may choose to distinguish as "syntactic" competence).

semantic theories for natural language. For if semantic competence is best characterized as the ability to understand the expressions of one's native I-language, then a theory of semantic competence is, in essence, a theory of linguistic understanding (LSC)—i.e., a theory of a particular aspect of a speaker's semantic psychology. In turn, if we take a semantic *theory* for natural language to be a theory of the objects of linguistic understanding, it follows that our best semantic theory can more or less double as a theory semantic competence.

8.2. Emphasizing the explanatory role of MSC

It has however been the central aim of this dissertation to demonstrate that a speaker's core semantic competence outruns the facts explained by current theories of LSC. Specifically I have argued that LSC needs to be supplemented by a theory of a speaker's metalinguistic-semantic competence (MSC). MSC is again so-called because by hypothesis it is grounded in a speaker's explicit (consciously accessible) conceptual knowledge of or beliefs *about* the meanings of linguistic expressions, and in particular the meanings of individual Words. Or to put the claim differently, I have argued that a speaker's explicit *conception of* a word's meaning is a direct reflection of his/her tacit understanding of the various ways in which the linguistically-encoded meanings of words guide and constrain without determining what those words can and cannot be *used/uttered* to talk about in ordinary discourse. To say that explicit knowledge of (or beliefs about, or conceptions of) lexical meanings is consciously *accessible* does not entail that such competence always comes to mind, as it were, during utterance interpretation. That is, such thoughts need not become *phenomenally conscious* in the traditional sense, but rather merely consciously available to language users upon

reflection. However, let me qualify this claim to allow for the possibility that very young children (perhaps even pre-linguistic infants) possess a degree of MSC which is conceptual in nature though perhaps not yet consciously available.

Now, given that MSC is grounded in extralinguistic conceptual knowledge, one might think that MSC belongs to a theory of CSC, properly understood. However, I have attempted to demonstrate that while fully conceptual in nature MSC is in fact a very close confederate of the non-conceptual resources that underwrite a speaker's LSC, and which non-linguistic animals lack. Indeed, I think of LSC and MSC, *collectively*, as constituting what I have been loosely calling a speaker's "core" semantic competence, which I define roughly as the ability to understand the meanings of expressions relative to a context of utterance.[3]

For instance, if what I argued in Chapter 5 is correct MSC often plays an important role in the contextual disambiguation of polysemous Words. And from this it follows that MSC is subsumable under the definition of semantic competence just proffered. In particular, I am suggesting that a theory of MSC is needed to fully *explain* how competent speakers understand the contextually-determined meanings of inherently polysemous Words. I have ruled out both CSC and PSC as possible *explanans* of such facts on grounds that such competence fall outside the realm of *linguistic* understanding, narrowly construed. Yet I again contend that a theory of I-languages/LSC cannot explain all the facts regarding linguistic understanding, and hence semantic competence.

[3] I *suspect* that a theory of MSC may even be needed to fully explain our interpretations of of "well-behaved" context-sensitive expressions such as indexicals and demonstratives, and indeed to represent the "character" (in Kaplan's sense) of any context-sensitive expression. However, I do not at the moment have a specific proposal and thus leave this question open for future investigation.

In addition to lexical polysemy, Chapter 4 was dedicated to illustrating how a theory of MSC can help explain how competent language users *acquire* Words with their lexical meanings. To be clear, I concede that MSC may not always be deployed in the acquisition of lexical meanings, as evidenced by the phenomenon of "fast mapping" whereby Word-learners are often able to immediately link unfamiliar Word-sounds to the concepts they express, or perhaps on just a few exposures.[4] In other cases, however, evidence suggests that Word-learning is a gradual process whereby learners often start out with a highly impoverished understanding of what an unfamiliar Word means in terms of which extralinguistic concept it is customarily used to express. In these circumstances, I have argued that learners use their concepts$_W$ of those Words to form metalinguistic beliefs about what they mean. With respect to a theory of semantic competence, I have noted that Chomskyans are committed to the idea that acquiring an I-language is part and parcel of one's core linguistic competence. And since the lexicon is widely considered to be an integral component of I-languages, broadly construed, it follows that the ability to acquire words with their meanings constitutes part of one's linguistic-semantic competence (LSC). Thus, MSC is needed to explain how learners acquire their lexicons. Moreover, if the acquisition of lexical meanings requires the exercise of one's MSC, then once again a theory of MSC is needed to complement a complete theory of core semantic competence, and by parallel reasoning any explanatorily adequate semantic theory for natural language.

My proposal in Chapter 7 is, admittedly, on a bit shakier ground in this regard. There I suggested that the role of MSC in the interpretation of identity statements is to

[4] Though recall discussion from Chapter 4 (section 4.2) on the development of metalinguistic-semantic competence.

recognize the information is *pragmatically conveyed* by an utterance of 'α is β', and specifically the metalinguistic fact that 'α' and 'β' are coreferential proper names (or again as Frege put it, "the circumstance that two names have the same content"). Thus, strictly speaking one might think that the competence employed in extracting this information from the linguistic context owes to what I have labeled *pragmatic-semantic competence* (PSC). However, even if recognition of the proposition that α is β corefer involves recruitment of extralinguistic pragmatic processes, understanding the information content of this proposition requires an understanding of what the names 'α' and 'β' *mean*. And if the meanings of names (or proper nouns) are fundamentally metalinguistic in nature, then to understand the meanings of names crucially involves the exercise of one's MSC.

8.3. Avenues for future research

It is one of my immediate future goals to close some of the gaps left open in the present study. In addition, I believe there is opportunity to more fully develop a theory of proper names whose semantics is grounded in a speaker's MSC—one that takes seriously the psychological facts about our understanding of the meanings of names. Indeed, I suspect that MSC may also role to play in our representation of the meanings of so-called "empty" names; i.e., names without bearers such as 'Pegasus'. I also think a theory of MSC has applicability to a theory of the semantics of so-called "quotation expressions."[5] More confidently, I feel quite sure that MSC has application to what Fiengo & May (2006) distinguish as *beliefs of Translation* (beliefs that non-coindexed (type-distinct), non-homophonous lexical expressions corefer).

[5] See Cappelen & Lepore (2007) for a thorough review of standing proposals.

In brief outline of the latter proposal,[6] consider that native Bostonians with thick accents pronounce the name 'Harvard' as something like \'hahvad\ and which upon first encounter can be quite difficult for Midwesterners such as myself to discern as meaning what its speakers are referring to, which of course is *Harvard University*. And so while non-native speakers may naturally hear the sound \'hahvad\ as a *possible* Word/name of natural language, it may not register in their idiolect (i.e., I-language) as a token of 'Harvard'. Thus, whereas the natives presumably lexicalize their singular concept HARVARD under the Word-sound \'hahvad\, non-natives may initially lexicalize the ontologically "empty" concept HAHVAD—say in the false belief that Harvard boasts a rival somewhere up the Charles called 'Havad University'. In other words, while the meaning of \'harverd\ is semantically associated with the non-native's concept HARVARD, \'hahvad\ is linked to HAHVAD. For such speakers, therefore, subsequent utterances of \'hahvad\ will be understood as an instruction to activate HAHVAD.

My proposal is that if/when misdirected speakers eventually come to realize that the sounds \'hahvad\ and \'harverd\ express the same concept, they will use their concept$_W$ of the Word 'Harvard' to form the *metalinguistic* belief that \'hahvad\ is a phonological variant of (i.e., *Translation* of) the Word-sound \'harverd\—i.e., that the sounds \'harverd\ and \'hahvad\ are phonological variants of the same name. Or put differently, we can that such speakers will come to believe that the Word-sounds \'harverd\ and \'hahvad\ "corefer."[7] In short, the ideas mentioned in this section are among those I plan to pursue further as part of my follow-on research.

[6] This topic was originally planned to occupy a chapter of the present study, but I ultimately decided to defer the project for further study.

[7] Though I find it plausible that over time such beliefs may become lexicalized as part of one's linguistic rather than purely metalinguistic competence.

8.4. Conclusion

Having made the best of the evidence available, I confess that this project wound up a bit overly ambitious in scope to fully flesh out the fine detail of each sub-thesis. And I am cognizant of other ways in which this thesis is lacking, particularly with respect to anticipating and responding to certain obvious objections to my view (along with certain difficulties that every language theorist is faced with). I am however confident that many such objections can, if pressed, be met, particularly if granted plausible background assumptions that I have helped myself to regarding the fundamental nature of language, thought, and the relationship between the two.

Shortcomings aside, I feel that among the most significant contributions of this dissertation is having been the first (so far as I know) to sketch a serious proposal regarding the psychological implementation of a speaker's MSC, which I remain steadfastly confident plays an important role in at least some forms of expression and/or utterance interpretation. Specifically, I hoped to have convinced readers that our concepts$_W$ of Words and the MSC they subserve provide ordinary speakers with a view, albeit dimly lit, onto the intrinsic semantic properties and relations of lexical items as manifest by their tokens in ordinary discourse. If I have succeeded to this extent, then my effort will have shed some new light on the nature of the relationship between language and thought, as well as the nature of semantic competence. My ultimate goal in this and work to follow is to motivate more cross-disciplinary research on the topic—research that might help either confirm or disconfirm the philosophically interesting/relevant aspects of the core thesis defended in these pages.

Bibliography

Akhtar, N., & Tomasello, M. (2001). The social nature of words and word learning. In R. M. Golinkoff, et. al., (eds.), *Becoming a word learner: A debate on lexical acquisition*, 115–135. Oxford: Oxford University Press.

Antony, L. M. & Davies, M. (1997). Meaning and Semantic Knowledge. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 71: 177-209.

Allwood, J. (2003). Meaning potentials and context: Some consequences for the analysis of variation in meaning. In Cuyckens, H., Dirven, R., & Taylor, J. R. (eds.), *Cognitive Approaches to Lexical Semantics*. Moulton de Gruyter, Berlin/New York.

Aydede, M. (2000). Computation and Intentional Psychology. *Dialogue*, 39, 4.

Bach, K. (1981). What's in a Name. *Australasian Journal of Philosophy*, 59: 371–86.

— (1987). *Thought and Reference*. Oxford University Press, Oxford.

— (2002). Giorgione was so-called because of his name. *Philosophical Perspectives*,16: 73–103.

Baker, M. C. (2003). *Lexical Categories: Verbs, Nouns and Adjectives*. Cambridge University Press, Cambridge.

Baldwin, D. A., and Moses, L. M. (1996). The ontogeny of social information gathering. *Child Development*, 67: 1915–1939.

Bar-Elli, G. (2006). Identity in Frege's *Begriffsschrift*: Where Both Thau-Caplan and Heck Are Wrong. *Canadian Journal of Philosophy*, 36, 3: 355-370.

Barner, D. & Bale, A. (2002). No nouns, no verbs: psycholinguistic arguments in favor of lexical underspecification. *Lingua*, 112: 771–791.

Beaney, M. (1997). *The Frege Reader*. Blackwell Publishers, Oxford.

Beretta, A. Fiorentino, R., and Poeppel, D. (2005). The Effects of Homonymy and Polysemy on Lexical Access: An MEG Study. *Cognitive Brain Research*, 24: 57– 65.

Bialystok, E. (2001). *Billingualism in development: Language, literacy, and cognition*. Cambridge University Press, New York.

Bierwisch, M. (1983). Semantische und konzeptuelle Repräsentation lexikalischer Einheiten". In Motsch, W. & Ruzicka, R. (eds.), *Untersuchungen zur Semantik*. Akademie Verlag, Berlin.

Bloom, P. (2000). *How Children Learn the Meanings of Words*. The MIT Press, Cambridge MA.

Bluntner, R. (1998). Lexical Pragmatics. *Journal of Semantics*, 15: 115-162.

Booij, G. E. (2007). *The Grammar of Words*. Oxford University Press, Oxford.

Borer, H. (2005a). *Structuring Sense. Volume I: In Name Only*. Oxford University Press, Oxford.

— (2005b): *Structuring Sense. Volume II: The Normal Course of Events*. Oxford University Press, Oxford.

Burge, T. (1973). Reference and proper names. *Journal of Philosophy: On Reference*, 70, 14: 425-439.

Cappelen, H. & Lepore, E. (2005). *Insensitive Semantics: A Defense of Semantic Minimalism and Speech Act Pluralism*. Blackwell Publishing.

— (2007). *Language Turned On Itself: The Semantics and Pragmatics of Metalinguistic Discourse*. Oxford University Press, New York.

Caramazza, A., & Grober, E. (1976). Polysemy and the structure of the subjective lexicon. In C. Rameh (ed.), *Georgetown University Round Table on Language and Linguistics*, 181-206, Washington, D.C.: Georgetown University Press.

Carey, S. (1978). The child as word learner. In Halle, M., Bresnan, J., & Miller, G. A. (eds.), *Linguistic theory and psychological reality*. The MIT Press, Cambridge MA.

Carey, S. & Bartlett, E. (1978). Acquiring a single new word. *Proceedings of the Stanford Child Language Conference*, 15: 17-29.

Carston, R. (1997). Enrichment and loosening: complementary processes in deriving the proposition expressed? *Linguistische Berichte*, 8: 103-127.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. The MIT Press, Cambridge MA.

— (1970): *Remarks on Nominalization*. Mouton & Co. N.V., Publishers, The Hague.

— (1981). *Lectures on Government and Binding: The Pisa Lectures*. De Gruyter, Berlin/New York.

— (1986). *Knowledge of Language: Its Nature, Origin, and Use*. Praeger Publishers, New York.

— (1995). *The Minimalist Program*. The MIT Press, Cambridge MA.

— (2000a). *New Horizons in the Study of Language and Mind*. Cambridge University Press, New York.

— (2000b). Minimalist Inquiries: The Framework. In Keyser, S. J., Martin, R., Uriagureka, J., Michaels, D. (eds.) *Step by Step. Essays on Minimalist Syntax in Honor of Howard Lasnik*. The MIT Press, Cambridge MA.

— (2003). Reply to Rey. In Antony, L. M. & Hornstein, N. (eds.), *Chomsky and His Critics*. Blackwell Publishing.

— (2005a). On Phases. In Freidin, R., Otero, C. P., and Zubizarreta, M. L. (eds.), *Foundational Issues in Linguistic Theory. Essays in Honor of Jean-Roger Vergnaud*. The MIT Press, Cambridge MA.

— (2005b). Three Factors in Language Design. *Linguistic Inquiry*. 36:1-22.

Clark, E. V. (1973). What's in a word? On the child's acquisition of semantics in his first language. In Moore, T. E. (ed.), *Cognitive development and the acquisition of language*. Academic Press, New York.

— (2009). Lexical Meaning. In Bavin, E. L. (ed.), The Cambridge Handbook of *Child Language*. Cambridge University Press, New York.

Clark, E. V. & Clark, H. H. (1979). When Nouns Surface as Verbs. *Language*, 55, 4: 767-811.

Copestake, A. & Briscoe, T. (1995). Semi-productive Polysemy and Sense Extension. *Journal of Semantics*, 12: 15-67.

Cruse, D. A. (2000). *Meaning in Language: An Introduction to Semantics and Pragmatics* (1st Edition). Oxford University Press, New York.

Dahan, D. & Magnuson, J. S. (2006). Spoken Word Recognition. In Traxler, M. J. & Gernsbacher, M. A. (eds.), *Handbook of Psycholinguistics* (2nd Edition). Elsevier.

Davidson, D. (1967). *The Logical Form of Action Sentences*. Reprinted in Davidson, D. (2001): *Essays on Actions and Events* (2nd Edition). Oxford University Press, Oxford.

— (1986). A Nice Derangement of Epitaphs. In Lepore, E. (ed.), *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*. Basil Blackwell, Oxford.

— (1990). The Structure and Content of Truth (The Dewey Lectures 1989), *Journal of Philosophy*, 87: 279–328.

Di Sciullo, A. M. & Williams, E. (1987). *On The Definition of Word*. Linguistic Inquiry Monographs 14. The MIT Press, Cambridge.

Dölling, J. (1995). Ontological Domains, Semantic Sorts and Systematic Ambiguity. *International Journal of Human-Computer Studies*, 43: 785-807.

Dummett, M. (1974). The Social Character of Meaning. *Synthese*, 27: 523-534.

— (1976). What is a Theory of Meaning? In Evans, G. & McDowell, J. (eds.), *Truth and Meaning*. Oxford University Press, Oxford.

Evans, G. (1981). Understanding Demonstratives. In Parret, H. & Bouveresse, J. (eds.) *Meaning and Understanding*. De Gruyter, Berlin/New York.

— (1982). *The Varieties of Reference*. Published posthumously, McDowell, J. (ed.). Oxford University Press, Oxford.

Fiengo, R. & May, R. (1998). Names and Expressions. *Journal of Philosophy*, 95: 377–409.

— (2006). *De Lingua Belief*. The MIT Press, Cambridge MA.

Fodor, J. (1975). *The Language of Thought*, Thomas Y. Crowell Company, Inc.

— (1998). *Concepts: Where Cognitive Science Went Wrong*. Oxford University Press, Oxford.

— (2008). *LOT-2: The Language of Thought Revisited*. Clarendon Press, Oxford.

— (2002). The Lexicon and the Laundromat. In Merlo, P. & Stevenson, S. (eds.), *The Lexical Basis of Sentence Processing: Formal, computational and*

experimental issues. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Fodor, J. & Lepore, E. (1996). The Pet Fish and the Red Herring: Why concepts aren't prototypes. *Cognition*, 58, 2: 243-276.

— (1998). The Emptiness of the Lexicon: Reflections on James Pustejovsky's *The Generative Lexicon*. *Linguistic Inquiry*, 29, 2: 269–288.

Frege, G. (various). In Beaney, M. (1997). *The Frege Reader*. Blackwell Publishers, Oxford.

— (1879). *Begriffsschrift* (*Concept Script*).

— (1884). *Die Grundlagen der Arithmetik* (*The Foundations of Arithmetic*).

— (1891). *Über Funktion und Begriff* (*On Function and Concept*).

— (1892). *Über Sinn und Bedeutung* (*On Sense and Reference*).

— (1893). *Grundgestze der Arithmetik*.

Geach, P. T., & Black, M. (1960, eds.). *Translations from the Philosophical Writings of Gottlob Frege*. Basil Blackwell, Oxford.

— (1962). *Reference and Generality: An Examination of Some Medieval and Modern Theories*. Cornell University Press, Ithaca NY.

Geurts, B (1997). Good news about the description theory of names. *Journal of Semantics*, 14: 319–348.

Gillette, J., Gleitman, L. R., Gleitman, H., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, 73: 135-176.

Gleitman, L. R. (1990). Structural sources of verb learning. *Language Acquisition*, 1: 1-63.

Goddard, C. (2000). Polysemy: A problem of definition. In Ravin, Y. & Leacock, C. (eds.), *Polysemy and Ambiguity: Theoretical and applied approaches*. Oxford University Press, Oxford.

Gorfein, D. S. (2002, ed.). *On the Consequences of Meaning Selection: Perspectives on Resolving Lexical Ambiguity*. Bluejacket Books/American Philosophical Association.

Grimshaw, J. (1979). Complement Selection and the Lexicon. *Linguistic Inquiry*, 10: 279-326.

— (1981). Subcategorization and grammatical relations. In Zaenen, A. (ed.), *Subjects and other subjects: Proceedings of the Harvard conference on the representation of grammatical relations*. Indiana University Linguistics Club, Bloomington IN.

Gupta, A. (1980). *The Logic of Common Nouns: An Investigation in Quantified Modal Logic*. Yale University Press, New Haven.

Hale, K. & Keyser, S. J. (1993). On Argument Structure and the Lexical Expression of Syntactic Relations. In Hale, K. & Keyser, S. J. (eds.), *The View from Building 20: Essays in Linguistics in Honor of Sylvain Bromberger*. The MIT Press, Cambridge MA.

— (2002). *Prolegomenon to a Theory of Argument Structure*. The MIT Press, Cambridge MA.

Halle, M., and Marantz, A. (1993): Distributed Morphology and the Pieces of Inflection. In Hale, K. & Keyser, S. J. (eds.), *The View from Building 20: Essays in Linguistics in Honor of Sylvain Bromberger*. The MIT Press, Cambridge MA.

Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 18: 441–461.

Harley, H. and Noyer, R. (1999): Distributed Morphology. *Glot International*, 4, 4: 3-9.

Harman, G. (1974). Meaning and semantics. In Munitz, M. K. & Unger, P. K. (eds.), *Semantics and Philosophy*. New York University Press, New York.

Heck, R. (2003). Frege on Identity and Identity-Statements: A Reply to Thau and Caplan. *Canadian Journal of Philosophy*, 33, 1: 83-102.

Heim, I., & Kratzer, A. (1998). *Semantics in Generative Grammar*. Blackwell Publishers, Malden, MA.

Higginbotham, J. (1988). Knowledge of Reference. In George, a. (ed.), *Reflections on Chomsky*. Oxford: Basil Blackwell.

Hintikka, J. (2004). *Analyses of Aristotle: Selected Papers*. Kluwer Academic Publishers, Dordrecht

Hornstein, N. (1986). *Logic as Grammar. An Approach to Meaning in Natural Language*. The MIT Press, Cambridge MA.

Horty, J. (2007). *Frege on Definitions: A Case Study of Semantic Content*. Oxford University Press, Oxford.

Izumi, Y. (2012). *The Semantics of Proper Names and other Bare Nominals*. PhD Dissertation, Department of Philosophy, University of Maryland, College Park.

Jackendoff (1999). What is a Concept, That a Person May Grasp It? In Murphy, G. L., (ed.), *Concept: Core Readings*. The MIT Press, Cambridge MA.

Julien, M. (2006). Words. In Brown, K. (ed.), *The Encyclopedia of Language and Linguistics* (2nd Edition), Elsevier.

Kamei, S., & Wakao. T. (1992). Metonymy: Reassessment, survey of acceptability and its treatment in machine translation systems. *Proceedings of ACL*, 1992, 309–311.

Kaplan, D. (1990). Words. *Proceedings of the Aristotelian Society*, 64: 93–119.

Karmiloff, K. & Karmiloff-Smith, A. (2001). *Pathways to Language*. Harvard University Press, Cambridge MA.

Katz, J. J. (1994). Names Without Bearers. The Philosophical Review, 103, 1: 1-39.

— (2001). The end of Millianism: Multiple bearers, improper names, and compositional meaning. *Journal of Philosophy*, 98, 3:137–166.

Klein, D. E., & Murphy, G. L. (2002). Paper has been my ruin: conceptual relations of polysemous senses. *Journal of Memory and Language*, 47: 548-570.

Klepousniotou, E. (2002). The Processing of Lexical Ambiguity: Homonymy and Polysemy in the Mental Lexicon. *Brain and Language*, 81: 205–223.

Klepousniotou, E., Titone, D., and Romero, C. (2008). Making Sense of Word Senses: The Comprehension of Polysemy Depends on Sense Overlap. *Journal of Psychology: Memory, Learning, and Cognition*. 34, 6: 1534–1543.

Kneale, W. (1962). Modality *de dicto* and *de re*. In Nagel, E., Suppes, P., & Tarski, A. (eds.), *Logic, methodology and philosophy of science. Proceedings of the 1960 International Congress*. Stanford University Press. Stanford, CA.

Konerding, K. P. (1997). Grundlagen einer linguistischen Schematheorie und ihr Einsatz in der Semantik. In Pohl, I. (ed.), *Methodologische Aspekte der Semantikforschung*. Lang, Frankfurt.

Koskela & Murphy (2006). Polysemy and Homonymy. In Brown, K. (ed.), *Encyclopedia of Language & Linguistics*, 2nd Edition, 2006: 724-744, Elsevier Ltd.

Kripke, S. (1979). A Puzzle about Belief. In Margalit, A. (ed.), *Meaning and Use*, 239-283. Dordrecht, D. Reidel.

— (1980). *Naming and Necessity*, Basil Blackwell, Oxford.

Landau, B., & Gleitman, L. R. (1985). *Language and experience: Evidence from the blind child*. Harvard University Press, Cambridge, MA.

Lany, J., & Saffran, J.R. (2010). From Statistics to Meaning, *Psychological Science*, 8 January 2010.

Larson, R. & Segal, G. (1995). Knowledge of Meaning. The MIT Press, Cambridge, MA.

Lepore, E. & Hawthorne, J. (2011). On Words. *Journal of Philosophy*, 108, 9: 447-485.

Levin, B. and Rappaport Hovav, M. (1998). Building Verb Meanings. In Butt, M. and Geuder, W. (eds.), *The Projection of Arguments: Lexical and Compositional Factors*. CSLI Publications, Stanford, CA.

Levin, B. and Rappaport Hovav, M. (2005). *Argument Realization*. Cambridge University Press, Cambridge.

Libben, G. & Jarema, G. (2007). Introduction: Matters of Definition and Core Perspectives. In Libben, G. and Jarema, G. (eds.), *The Mental Lexicon: Core Perspectives*. Elsevier, Amsterdam.

Ludlow, P. (2011). *The Philosophy of Generative Linguistics*. Oxford University Press, Oxford.

Lust, B. (2006). *Child Language: Acquisition and Growth*. Cambridge University Press, Cambridge.

Macnamara, J. (1982). *Names for things. Cambridge*, MA: MIT Press.

Mampe, B., Friederici, A.D., Christophe, A., & Wermke, K. (2009) Newborns' Cry Melody Is Shaped by Their Native Language. *Current Biology*, 19, 23: 1994-1997.

Margolis, E. & Laurence, S. (1999). Concepts and Cognitive Science. In Margolis, E. & Laurence, S. (eds.), *Concepts: Core Readings*. The MIT Press, Cambridge MA.

Markman, E. M. (1994). Constraints children place on word meanings. In Bloom, P. (ed.), *Language acquisition*. The MIT Press, Cambridge MA.

May, R. (2000). Frege on Identity Statements. In Cecchetto, C., Chierchia, G., & Guasti, M. T. (eds.), *Semantic Interfaces: Reference, Anaphora and Aspect,* Stanford, CSLI Publications. (Citations here are from an earlier draft).

McCawley, J. D. (1968). The role of semantics in a grammar. In Bach, E. and Harms, R. T. (eds.), *Universals in linguistic theory*. Holt, Rinehart and Winston, New York.

Murphy, G. L. (2002). *The Big Book of Concepts*. The MIT Press, Cambridge MA.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92: 289-316.

Murphy, M. L. (2003). *Semantic Relations and the Lexicon*. Cambridge University Press, Cambridge.

Nunberg, G. (1979). The non-uniqueness of semantic solutions: polysemy. *Linguistics and Philosophy*, 3.1, 1979.

Ostertag, G. (2007). Review of Fiengo & May (2006): *De Lingua Belief*. *Notre Dame Philosophical Reviews*.

Paul, I. & Stainton, R. (2009). Review of Wolfram Hinzen's: *An Essay on Names and Truth*. *Mind*, 118: 472-475.

Perry, J. (2003). *Frege on Identity, Cognitive Value, and Subject Matter*. PhilPapers: http://philpapers.org.

Pesetsky, D. (1982): *Paths and Categories*. Doctoral dissertation, Massachusetts Institute of Technology, Department of Linguistics and Philosophy.

Pethõ, G. (2001).What is Polysemy? A Survey of Current Research and Results. In Enikő Németh, E., & Bibok, T. K. (eds.), *Pragmatics and the Flexibility of Word Meaning*. Elsevier.

Pietroski, P. M. (2000). The undeflated domain of semantics. *The Nordic Journal of Philosophy*,1: 161–76.

— (2005a). *Events and Semantic Architecture*. Oxford University Press.

— (2005b). Meaning before truth. In Preyer, G. & Peter, G. (eds.), *Contextualism in Philosophy: Knowledge, Meaning, and Truth*. Oxford University Press, Oxford.

— (2006a). Character Before Content. In Thomson, J. & Byrne, A. (eds.), *Content and Modality: Themes from the Philosophy of Robert Stalnaker*, Oxford University Press, New York.

— (2006b). Interpreting concatenation and concatenates. *Philosophical Issues*,16: 221–245.

— (2007). Systematicity via Monadicity. *The Croatian Journal of Philosophy*, 7: 343–374.

— (2008). Minimalist Meaning, Internalist Interpretation. *Biolinguistics*, 2, 4: 317–341.

— (forthcoming): *Semantics Without Truth Values*. Oxford University Press, Oxford.

Pinker, S. (1994). *The Language Instinct: How the Mind Creates Language.* New York: HarperCollins.

Pustejovsky, J. (1991). The Generative Lexicon. *Computational Linguistics*, 17, 4: 409-441.

— (1995). *The Generative Lexicon*. The MIT Press, Cambridge, MA.

— (1998). Generativity and Explanation in Semantics: A Reply to Fodor and Lepore. *Linguistic Inquiry*, 29, 2: 289–311.

— (2005). Generative Lexicon. In Brown, K. (ed.), *The Encyclopedia of Language and Linguistics* (2nd Edition), Elsevier.

Pylkkänen, L., .Rodolfo, L., & Murphy, G. L. (2006). The Representation of Polysemy: MEG Evidence. *Journal of Cognitive Neuroscience*, 18, 1: 97–109.

Pylyshyn, Z. (2003). Seeing and Visualizing: It's Not What You Think. Cambridge, MA: The MIT Press.

— (2009). Perception, Representation and the World: The FINST that binds. In Dedrick, D. & Trick, L. M. (eds.), *Computation, Cognition and Pylyshyn*. Cambridge, MA: The MIT Press.

Quine, W. V. O. (1960). *Word and Object*. The MIT Press, Cambridge MA.

Radford, A. (2009). *Analysing English Sentences: A Minimalist Approach*. Cambridge: Cambridge University Press.

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

Recanati, F. (2004). *Literal Meaning*, Cambridge: Cambridge University Press.

Rey, G. (1985): *Concepts and Conceptions: A Reply to Smith, Medin, and Rips*. Cognition, 19: 297-303.

Rodd, J., Gaskell, G., and Marslen-Wilson, W. (2002): *Making Sense of Semantic Ambiguity: Semantic Competition in Lexical Access*. Journal of Memory and Language, 46: 245–266.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7: 573–605.

Rubio-Fernandez, P. (2008). *Concept Narrowing: The Role of Context-independent Information*. *Journal of Semantics*, 25: 381–409.

Ruhl, C. (1989). *On Monosemy: A Study in Linguistic Semantics*. State University of New York Press, Albany.

Russell, B. (1903). *The Principles of Mathematics*. Norton, New York, New York, 2nd edition, 1938.

— (1905). On Denoting. *Mind*. 14, 4: 479–493.

— (1972). *The Philosophy of Logical Atomism*. Open Court Publishing.

Salmon, N. (1986). *Frege's Puzzle*. The MIT Press, Cambridge, MA.

Saul, J. (1997). Substitution and Simple Sentences. *Analysis*, 57: 102–8.

— (2007). *Simple Sentences, Substitution, and Intuitions*. Oxford University Press, Oxford.

Segal, G. (2001). Two Theories of Names. *Mind & Language*, 16, 5: 547–563.

Sharwood-Smith, M. (2004). In two minds about grammar: On the interaction of linguistic and metalinguistic knowledge in performance. *Transactions of the Philological Society*, 102, 3: 255-280.

— (2010). Metalinguistic processing and acquisition within the MOGUL framework. In De Mulder, H., M. Everaert, Ø. Nilsen, T. Lentz & A. Zondervan (eds.), *Theoretical Validity and Psychological Reality*. John Benjamins Publishing, Amsterdam.

Sluga, H. (1986). Semantic Content and Cognitive Sense. In Haaparanta, L. &. Hintikka, J. (eds.), *Frege Synthesized: Essays on the Philosophical and Foundational Work of Gottlob Frege* (Synthese Library). Dordrecht, Netherlands.

Soames, S. (2002). *Beyond Rigidity: The Unfinished Semantic Agenda of Naming and Necessity*. Oxford University Press, New York.

Soler, C. and M. A. Marti (1993). *Dealing with Lexical Mismatches*, Esprit BRA-7315 Acquilex2 Working Paper. 4.

Smith, B. (1992). Understanding Language. *Proceedings of the Aristotelian Society* 92: 109-141.

— (2006). What I Know When I Know a Language. In Lepore, E. & Smith, B. (eds.), *The Oxford Handbook of Philosophy of Language*, Oxford University Press, Oxford.

Smith, E. E., & Medin, D. M. (1981). *Categories and concepts*. Harvard University Press, Cambridge, MA.

Smith, E. E., & Osherson, D. (1984). Conceptual combination with prototype concepts. *Cognitive Science*, 8: 337-361.

Snedeker, J. (2009). Word Learning. In Squire, L. (ed.), *Encyclopedia of Neuroscience*. Academic Press.

Sosa, D. (1996). The Import of the Puzzle About Belief. *The Philosophical Review*, 105, 3: 373-402.

Sperber, D. & Wilson, D. (1986). *Relevance: Communication and Cognition*. Oxford: Blackwell (2nd Edition, 1995).

Srinivasan, M. & Snedeker, J. (2011). Judging a book by its cover and its contents: The representation of polysemous and homophonous meanings in four-year-old children. *Cognitive Psychology*, 62: 245–272.

Strawson, P. F. (1959). *Individuals*. Methuen, London.

— (1974). *Subject and Predicate in Logic and Grammar*. Methuen, London.

Swingley, D. (2010). Fast Mapping and Slow Mapping in Children's Word Learning, *Language Learning and Development*, 6: 179–183.

Taylor, J. R. (2003). Cognitive models of polysemy. In Nerlich, B., Todd, Z., Herman, V., & Clarke, D. D. (eds.), Polysemy. Mouton De Gruyter, Berlin.

Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.

Thau, M. & Caplan, B. (2003). What's Puzzling Gottlob Frege? *Canadian Journal of Philosophy*, 31: 159-200.

Williams, E. (1994). *Thematic Structure in Syntax*. The MIT Press, Cambridge, MA.

— (2007). Dumping Lexicalism. In Ramchand, G. & Reiss, C. (eds.), *The Oxford Handbook of Linguistic Interfaces*. Oxford University Press, Oxford.

Woodward, A. (2004). Infants' Use of Action Knowledge to Get a Grasp on Words. In Waxman, S., and Hall, D. G. (eds.), *Weaving a Lexicon*, The MIT Press, Cambridge, MA.

Wunderlich, D. (2006). Introduction: What the Theory of the Lexicon is About. In Wunderlich, D. (ed.), *Advances in the Theory of the Lexicon*. Mouton De Gruyter, Berlin.

Zalta, E. N. (2001). Fregean Senses, Modes of Presentation, and Concepts. *Philosophical Perspectives*, 15: 335-359.