Natural Language and Formal Languages

by

Josep Macià Fàbrega

*Llicenciat en Filosofia*, Universitat de Barcelona (1988)

Submitted to the Department of Linguistics and Philosophy
in Partial Fulfilment of the Requirements for the Degree of

Doctor of Philosophy in Philosophy

at the

Massachusetts Institute of Technology

June 1997

© 1997 Josep Macià Fàbrega. All Rights Reserved

The author hereby grants to MIT permission to reproduce and to
distribute publicly paper and electronic copies of this thesis document in whole or in part.

Signature of Author .......................... .......................... ..........................

Department of Linguistics and Philosophy
November 25, 1996

Certified by ..................................

Robert Stalnaker
Laurance S. Rockefeller Professor of Philosophy
Thesis Supervisor

Accepted by ..........................

Alex Byrne
Assistant Professor of Philosophy
Chairman, Departmental Committee on Graduate Studies

JUN 12 1997
ARCHIVES
LIBRARIES
Natural Language and Formal Languages

by

Josep Macià Fàbrega

Submitted to the Department of Linguistics and Philosophy
on November 27, 1996 in Partial Fulfilment of the Requirements
for the Degree of Doctor of Philosophy in Philosophy

Abstract

This thesis consists of three papers on the semantics of natural language and formal languages. Chapter one discusses how the possible interpretation of the noun phrases in a sentence is affected by the syntactic structure of the sentence. In particular, we focus on the phenomena related to principles (B) and (C) of binding theory. We can explain all these phenomena, including the counterexamples that have been offered against standard binding theory, by (i) viewing the binding principles as semantic principles that constrain the relation of presupposed co-reference (ii) correctly identifying the context that is relevant for the application of the binding principles, and (iii) viewing binding theory as just one among several sources of semantic information that are used when interpreting the noun phrases in a sentence.

Chapter two attempts to provide support for George Boolos' proposal of interpreting second-order quantification by means of plural quantification by arguing that several of the criticisms that have been offered against it are not correct. We consider the criticisms put forward by Charles Parsons, Patrick Grim and Michael Resnik.

Chapter three distinguishes two main senses in which a formal language can be said to have been provided with an interpretation: (i) by having provided a model (or structure) and a definition of satisfaction and truth in the standard way (ii) by having provided a translation into a natural language. We argue that the sentences of a formal language interpreted as in (i) do not mean anything. A formal language interpreted as in (i) models the way the truth of a sentence would be affected by two factors: the interpretation as in (ii) of the language, and a way the world might be. Viewing in this way the relation between interpreting a formal language as in (i) and as in (ii) allows us to justify the conceptual adequacy of the standard model-theoretic definitions of the properties of logical truth and logical consequence.

Thesis supervisor: Robert Stalnaker
Title: Laurance S. Rockefeller Professor of Philosophy
This dissertation is
dedicated to my mother,
Maria de l’Esperança Fàbrega i Gregori
Acknowledgements

George Boolos was my advisor since I first got to MIT. He supervised my work up to the last stages of this dissertation. I owe him much. I enjoyed seeing him in action in my many courses with him. Solving his problem sets was lots of fun and suffering -- simultaneously! He was always, always, willing to devote his time to me whenever I wanted to talk to him. I am very grateful for this. George had a pretty relaxed style of tutoring. If he had put more pressure on me -- actually, if he had put any pressure on me -- I might have finished this dissertation much earlier. I don’t regret that at all, though. It would not have been this dissertation. It was a privilege to be able to enjoy his wit, his splendid mind and his humor. I loved him and I miss him.

I feel gratitude and admiration for the other two members of my original thesis committee: Bob Stalnaker and Irene Heim. Through his writings, his lectures and many conversations Bob has strongly influenced my ideas in philosophy. Talking to him would always help me see the more general issues that laid behind any specific discussion in this dissertation. From Irene I have learned a lot. Her long written comments on my answers to her problem sets made me think and try to look deeper into the topics discussed. I can only wish to emulate her as a teacher. She read with interest parts of the dissertation that might not have seemed related to the areas she is working on, and provided me with important comments. Her views on binding theory have had an obvious influence on the approach I defend in chapter one.

I thank Vann McGee for kindly accepting to join my thesis committee in the final stages, for providing some helpful comments, and for forcing me not to take for granted any of the claims I wanted to defend in chapter three.

I will always be indebted to all the other teachers I had during my time at MIT. I could write long about the special way in which each one has influenced me. I will just write their names: Dick Cartwright, Judy Thomson, Sylvain Bromberger, Josh Cohen, Ned Block, Paul Horwich, Jim Higginbotham, Noam Chomsky, Alec Marrantz and David Pesetsky.

I believe I have been lucky to be at the Linguistics and Philosophy department at MIT at a very special time. Besides the people I already mentioned there are many others that helped make this place what it was during my stay here. I have benefited from their friendship, their being my classmates or fellow participants in reading groups, their hospitality, their linguistic judgements, and many, many discussions. They include: Gabriel Uzquiano, Joe Lau, Jason Stanley, Kathrin Koslicky, Daniel Stoljar, Robert Kermode, Joana Rosselló, Darryl Jung, David Hunter, Lenny Clapp, Rob Stainton, Brian Ulicny, Tim Dilligan, Michael Glanzberg, Richard Heck, Alex Byrne, Ned Hall, Chris Collins, Eulàlia Bonet, Sabine Iatridou, Chris Tancredi, Utpal Lahiri, Friederike Moltmann, Pilar Barbosa, Diana Cresti, Albert Branchadell, Pierre Pica, Lenny Katz, John Pyrovolakis, Zoltán Szabó, Tim Hinton, Tony Gray, Jennifer Noonan, Aviv Hoffmann, Stephen Walther, Lisa Sereno, Cara Spencer, Judy Feldmann, Andrew Botterell, Matti Eklund,

I would not have been able to come to graduate school at MIT if I had not been lucky enough to have some good teachers and some extraordinary fellow students as an undergraduate at the University of Barcelona. I wish to thank them all. I wish to thank specially Manuel García-Carpintero and Ignasi Jané. I have had very helpful discussions with each of them on some of the topics in these dissertation.

During my first two years at MIT I was supported by a grant from 'La Caixa'. MIT and the Linguistics and Philosophy department supported me for two additional years, and will support as well the final filing of the thesis. I am grateful for this financial support. I hope, though, that it will be possible to improve the funding situation for future students.

In addition to the members of my thesis committee, several other people provided me with comments on parts of this dissertation; among them: Joana Rosselló, Alec Marantz, Joe Lau, Gemma Rigau, Howard Lasnik, Jim Higginbotham, Ignasi Jané, Manuel García-Carpintero, Matti Eklund, Michael Glanzberg, Manuel Perez, Gabriel Uzquiano and Richard Cartwright.

Wayne O'Neil, the head of the department, has always been ready to assist me with a cheerful mood, which is much appreciated. I thank Jen Purdy for being so friendly and efficient, and making the dealing with burocratic matters much easier than it would otherwise have been, particularly in the past year while I was in Barcelona.

During the five years I lived in Cambridge it was always a source of confort to read Ambrós Domingo‘s long letters, and to conversate with James Blanchflower (both p and not p). I’m grateful to Merche Fages for her friendship and hospitality.

Above all I would like to thank my wife, Susanna Pi, for her patience, criticism, generosity and some one thousand and ten reasons more. And I am also grateful to little Cinta for helping me to realize that writing a dissertation and getting a Ph.D. is, after all, something of very little importance...
Natural Language and Formal Languages

Table of Contents

I. Abstract ..................................................... 3

II. Acknowledgements ............................................. 6

III. Table of Contents .............................................. 9

IV. Chapter One
    Sentence Structure, Noun Phrase Interpretation
    and Context ..................................................... 11

V. Chapter Two
    On the Interpretation of Second-order Quantification:
    An Examination of Some Criticisms of Boolos’ Proposal .......... 87

VI. Chapter Three
    On the Interpretation of Formal Languages and
    the Analysis of Logical Properties ................................ 143
0. Introduction

Structural relations among Noun Phrases (or NPs) in a sentence place constraints on the possible semantic relations among those NPs. For instance, in (1a)

(1) a. Andrew saw him
   b. Everybody blamed him

Andrew and him can not refer to the same individual, and (1b) can not mean that everybody blamed himself, that is, him can not be interpreted as bound by Everybody --'bound' in the sense analogous to a logical variable being bound by a quantifier. Binding Theory attempts to account for facts like the ones just described.

In this paper we focus on the phenomena related to the so called Principles (B) and (C) of the Binding Theory and specially on the different kinds of sentences that have been offered as counter-examples. I contend that we can account for the phenomena by: (i) Viewing Binding
Theory as a semantic theory about how to semantically interpret certain structural relationships among NPs. (ii) Modifying principles (B) and (C) so that they are not about the relation of having the same value, but rather about the relation of it being presupposed to have the same value in a given context. (iii) Identifying what should be regarded as the relevant context for the application of the Binding Principles. (iv) Identifying how the Binding Principles interact with other sources of semantic information in order to yield the interpretation for a sentence.

The paper has eight sections. In section 1 I present what I will call Standard Binding Theory. In section 2 I present some counter-examples for Standard Binding Theory that have been offered in the literature. In section 3, I give an overview of the proposal I will defend and, following Heim(1993), I give an alternative formulation of principles (B) and (C). In section 4 I show how we can explain the acceptability of some of the alleged counter-examples: those crucially involving the interaction of different sources of semantic information. In section 5.1 we show how under the Binding Principles we propose we can account for the counter-examples involving 'non-presupposed coreference'. In section 5.2 we focus on another group of alleged counter-examples: identity sentences. I argue that one attempt of dealing with these sentences by appealing to pragmatic considerations is not satisfactory; I further content that there are syntactic reasons for thinking that, nevertheless, identity sentences do not really pose a problem for Binding Theory. In section 6 we make some heterogeneous remarks and further develop some points made in the previous sections. In section 7 we discuss the necessity of distinguishing two different stages in the interpretation of a sentence. Finally, in section 8 I present another kind of sentences that would seem to pose a problem both for Standard Binding Theory and for the
version of Binding Theory that I defend. As far as I know, this kind of sentences has not been considered before in the literature. I will defend that they can be accounted for by correctly determining what the relevant context for the application of the binding principles is.

1. **Standard Binding Theory**

1.1. **Principles (A), (B) and (C)**

Consider the following examples, where two NPs in a sentence have the same index if and only if either they both refer to the same individual, or one of them is a quantificational NP (like, e.g, *Every boy*, or *Somebody*) and the other is bound to it; a "*" indicates that the sentence (with the specific interpretation indicated by the indexation) is not acceptable:

(2)  

a. Cara, saw her,  

b. *Cara, saw her,  

c. She, saw her, father  

d. Matti, says that he, is tired  

e. Matti, says that he, is tired  

f. He, says that Tony, is tired  

g. *He, says that Tony, is tired
h. His father says that Samuel is tired

i. *(His father) says that Robert is tired

j. (His father) says that he is tired

k. He saw Dean

l. *He saw Dean

m. His mother saw Dean

n. *Everybody saw him

o. Everybody saw him

p. Every thief fears that he gets caught

q. In Tarragona, every single boy’s mother sends the twerp off to summer camp

r. *In Tarragona, every boy thought the twerp would hate summer camp

We can account for all the data in (2) by appealing to conditions such as Chomsky’s Principles (B) and (C) of the Binding Theory. We will use the term Standard Binding Theory to refer to the Binding Theory introduced in Chomsky(1981). In this section we will present Standard Binding Theory. Standard Binding Theory has three principles: the so called principles (A), (B) and (C). In this paper we will be concerned only with the phenomena related to principles (B) and (C). Nevertheless, for completeness in this first section we will introduce principle (A) along with principles (B) and (C). In order to explain what these principles are we need first to introduce the relation of c-commanding.
The relation of *c-commanding* is a diadic relation between nodes in a tree. It can be specified by means of the following definitions:

Node *X immediately dominates* node *Y* if *Y* is a daughter of *X* (i.e., if *X* is higher up in the tree than *Y* and *Y* is connected to *X* by a single branch).

Node *X dominates* node *Y*, if either *X* immediately dominates *Y*, or there is a *Z* such that *X* dominates *Z* and *Z* immediately dominates *Y*.

Node *X c-commands* node *Y* if any branching node dominating *X* also dominates *Y*, and *X* itself does not dominate *Y*, nor *Y* dominates *X*.

For instance in (3a), node *B* c-commands *C* and *E* (but not *A* or *D*), *D* c-commands the same nodes as *B*; *A* does not c-command any node; *C* c-commands *B* and *D*, and so does *E*. Notice that in (3b), unlike what is the case in (3a), *D* does not c-command *C* or *E*.

---

1 If we define a tree as a non-empty finite sequence, and we inductively define 'N is a node of tree T' as:

\[ \exists X_1,...,X_m (T=<X_1,...,X_m> \text{ and } (\exists i (1\leq i \leq m \text{ and } (N=X_i \text{ or } N \text{ is a node of } X_i))) \]

then we can define 'node X immediately dominates Y' as:

\[ \exists X_1,...,X_m (X=<X_1,...,X_m> \text{ and } (\exists i (1\leq i \leq m \text{ and } Y=X_i))) \]
Consider now the following sentences

(3)  

a.  [Zoltan's brother]i saw himselfi  

b.  Zoltani saw himselfi  

c.  *Zoltani’s brother saw himselfi  

d.  *Zoltani thinks that Mary saw himselfi  

The contrast between (3a-b) and (3c) illustrates that a reflexive pronoun has to have an antecedent that c-commands it. We can see this more clearly by considering the syntactic structure of the sentences. The (simplified) syntactic representations of (3a) and (3c) are, respectively, (4a) and (4b)
The whole NP Zoltan's brother c-commands himself, but the NP Zoltan does not. The former, then, can be an antecedent of himself, whereas the latter can not. On the other hand, the contrast between (3b) and (3d) illustrates that the antecedent must be 'close enough' to the reflexive.

The dada in (3) is accounted for by the Principle (A) of Standard Binding Theory

**Principle A:** An anaphor must be bound in its Governing Category.

Regarding the terminology employed: An anaphor is a reflexive pronoun or a reciprocal; an NP Y is bound by an NP X if X c-commands Y, and X and Y are coindexed; the notion of Governing Category purports to capture the idea of 'being close enough' that we mentioned above. For the purposes of the present discussion we can set aside some difficulties arising when trying to specify the notion of Governing Category, and we can take the Governing Category (GC) of a node X to be the smallest (i.e. the one containing less nodes) NP or S that contains X. So, in (3a-c) the GC for the reflexive is the whole sentence, in (3d) is the Complementizer Phrase (CP) that Mary saw himself. In (5b) below, the GC for him is the whole sentence, whereas in (5)d is the CP that she loves him, and in (5)e the GC for his is the NP his mother.

(5) a. *Gabriel, saw him_i
   b. Gabriel, saw him_j
   c. His, mother loves him_i
   d. Gabriel, thinks that she loves him_i

-17-
In the sentences in (5) we can observe the following: a non-reflexive pronoun (also called a pronominal) does not need to be bound, as illustrated by the pronominal him in (5b) and (5c) (notice that in (5c) His does not c-command him, even though His mother does); a pronominal does not need to be bound but it can be bound, as illustrated in (5d) and (5e); a pronominal, though, can not be bound in its governing category as (5a) illustrates. All these observations follow from the Principle (B) of the Standard Binding Theory:

**Principle (B):** A pronominal must be free in its Governing Category

A node is **free** if it is not bound by any other node.

The examples in (6) show that names and definite descriptions are subject to constrains stronger than those for pronominals

(6) a. *She, saw Rachel,

b. *She, thinks that Rachel, is smart

c. *Rachel, thinks that he loves Rachel,

d. *She, thinks that [the woman with the blue hat], is smart

The sentences in (6) suggest that a proper name or a definite description must be free not
only in its GC but everywhere in a sentence. For instance, in (6b) She binds Rachel, and the structure is unacceptable even though She is not in the GC for Rachel, which is the CP that Rachel is smart.

If we use coindexation to also represent the relation between a quantificational NP and the NPs it binds ('binding' not in the syntactic sense we have just introduced but in the logical sense of a logic variable being bound by a quantifier), we observe that all the constraints affecting a proper name or definite description also affect a quantificational NP:

(7)  

a. *He saw [some man],

b. *He thinks that [every man], is smart

c. *[Some man], thinks that Lisa loves [some man],

d. *He thinks that [some man], is smart

Proper names, definite descriptions and quantificational NPs are grouped under the category of \textit{R-expressions}. Principle (C) constrains their distribution. It completely agrees with

\footnote{Quantificational NPs are actually subject to stronger constraints than proper names and definite descriptions. We can see this in the contrast between (i) and (ii)}

(i) His mother thinks that Gabriel, is smart

(ii) *His mother thinks that every man, is smart

In both (i) and (ii) none of the two NPs binds the other. Binding Theory does not rule out either sentence. Several principles independent of Binding Theory have been proposed that would account for the ungrammaticality of (ii), in particular the so called bijection principle (Koopman and Sportiche(1982) p.140)
what we have observed with respect to the examples in (6) and (7):

**Principle (C):** R-expressions must be free (everywhere).

As mentioned above in this paper we will concern ourselves only with the phenomena related to binding principles (B) and (C).

Different sets of conditions can be formulated that, even if they are not equivalent to (B) and (C), are similar to them and are intended to account for the same kind of phenomena. We will call *disjointness conditions* to any such set of conditions.

1.2. *Some Comments*

Notice that we are not using indexes simply as part of some provisional notational convention to help us indicate what the reading of a sentence that we want to consider is. Rather, in the way we are using them, indexes are an essential part of the syntactic representation of a sentence (in the same way that each node in a tree is part of that representation). This is so because the Binding Principles are formulated in terms of the relation of *binding*, which, in turn, appeals to the relation of *sameness of indexes*.

Also notice that in order to account for the data in (1)-(7) we need to appeal not only to the Binding Principles but also to a principle such as "Referential NPs with the same index must
refer to the same individual, referential NPs with different indexes must refer to different individuals; a NP is interpreted as bound by a quantificational NP if and only if they have the same index. We mentioned this principle above as in passing, but notice that it is crucial. Without it, Binding Theory would not explain, for instance, why (8a)

(8) a. Lisa likes her
b. Lisaᵢ likes herᵢ
c. Lisaᵢ likes herᵢ
d. Lisaᵢ likes herᵢ
e. Lisaᵢ likes herᵢ

is unacceptable if Lisa is coreferential with her: Even if (8b) is ruled out by Principle (B), (8c) is not. If it were possible for two NPs with different indexes to have the same referent, then (8c) could give rise to an interpretation that (8a) does not have.

This discussion shows also another point: a theory that tries to explain why (8a) can not have a coreferential reading has to include, some way or other, a semantic component. The explanation can not be given in purely syntactic terms. It might be, as it is the case with Standard Binding Theory, that the bulk of the explanation is carried out by purely syntactic principles, and that the semantic component is straightforward. Still the semantic part, even if straightforward, is essential to explain what we want to explain.

It is possible to have a theory where the bulk of the explanation is at the semantic level. We can, for instance, formulate disjointness conditions which are completely analogous to
Standard Binding Theory principles (B) and (C) but that operate at the semantic level (that is, they are conditions that directly constrain which interpretations are possible for a given syntactic structure). Chomsky (1993) offers one set of such a semantic kind of disjointness conditions. Principles (B) and (C) are formulated there, roughly, in the following way:

(B) If $\alpha$ is a pronominal, interpret it as distinct in reference from every c-commanding phrase in its GC.

(C) If $\alpha$ is an $r$-expression, interpret it as distinct in reference from every c-commanding phrase.

(In order to account for cases like (2n)-(2r) where there is not co-reference involved, these principles should, of course, be modified --or at least the term "distinct reference" should be re-defined in such a way that it also applies to a relation such as the one between Everybody and him in (2n). With such a modification in place, this theory makes the same predictions as Standard Binding Theory with respect to what interpretations the different sentences can have).

2. Counter-examples to Standard Binding Theory Principles (B) and (C)

It has been observed in the literature that there are several different kinds of sentences that are counter-examples to the Standard Binding Theory Principles (B) and (C). In (9) we have some
examples of sentences that pose a problem for principle (C)³:

(9)  a. Who is that woman over there? She is professor Rigau

b. Everyone has finally realized that Oscar is incompetent. Even he has finally realized that Oscar is incompetent.

c. Mary, Betty and John have one thing in common: Mary admires John, Betty admires John, and he himself also admires John.

d. What do you mean Oscar loves no one? He loves Oscar

e. The logic tutor while trying to explain the law of Universal Instantiation to a student tells him: Look fathead. If everyone loves Oscar, then certainly he himself must love Oscar

f. I think that woman talking on TV is Zelda. She says the same things that Zelda says in her book

Each of these sentences can be interpreted so that the two italicized NPs have the same referent.

³ Even if I do not specify the source of each of these sentences, most of them are adapted or directly taken from the references mentioned at the end; the same is true of the sentences in (10).
If they are interpreted as having the same referent, then the two NPs must have the same index in the syntactic structure of the sentence. In each of these sentences, though, one NP c-commands the other, and, furthermore, the c-commanded NP is an R-expression. The syntactic structure, so, is undesirably ruled out by Principle (C) of the Standard Binding Theory.

Analogously, the sentences in (10) pose a problem for the Principle (B) of the Standard Binding Theory

(10) a. Mary, Betty and John have one thing in common: Mary admires John, Betty admires him, and John himself also admires him.

b. At a Halloween party someone says: I think that the man with the devil costume is Joe. It is suspicious that he knows him so well.

c. What do you mean no one loves Oscar? Oscar loves him

d. What do you mean no one loves Oscar? He loves him

Some of these sentences (like (9a) or (10b)) are one hundred per cent acceptable, others are not perfect but are good enough as to pose a problem for a theory according to which they are simply ruled out.
Heim(1993) offers some examples involving quantification which she claims are marginally acceptable and, to the extent that they are acceptable, also pose a problem for Principle (B), as for instance

(11) a. The logic tutor while trying to explain the law of Universal Instantiation tells his student: Look, if everyone loves Oscar then it surely follows that Oscar himself must love Oscar. And, of course, this does not hold just for Oscar, but for any arbitrary man: if everyone loves a man, then that man himself must love him

b. Somebody said that what he had in common with his siblings was that his sister admired him, his brother admired him, and he himself also admired him

3. An alternative proposal

3.1. Other Proposals

There are several ways in which we could try to deal with these problematic examples: One possibility would be to claim that the disjointness conditions should not be constraints on the relation of 'co-indexation' or of 'co-reference' but rather they should constrain some other relation, and that when this is done the disjointness conditions will allow just those interpretations that the sentences can actually have. This is the approach taken in Higginbotham(1992). Of course the difficult task when opting for this view is to specify a suitable relation in terms of which we can formulate the disjointness conditions. Higginbotham introduces the relation of
common reference, and formulates the disjointness conditions in terms of it. This relation is
defined in terms of the intentions of the speaker.

Another possible approach is to accept principles (B) and (C) as they are in Standard
Binding Theory, but to modify the semantic part of the theory. Recall that the principle on the
interpretation of indexes requires that two NPs get assigned the same individual if and only if
they have the same index (and it also constrains how to interpret those NPs that will have an
interpretation analogous to bound variables). One possible modification to the principle is to
require only that if two NPs have the same index then they get assigned the same individual, but
allowing that NPs with different indexes might also get assigned the same individual.
Fiengo&May(1994) takes this kind of approach. The difficult part for this approach is not to
explain, for instance, why (9b) can have a co-referential interpretation, but rather why (8a) can
not usually have it. Fiengo&May(1994) tries to solve this difficulty by appealing to pragmatic
principles.

Another possibility is to think that in addition to disjointness conditions, we should also
have some other principle which will allow us to make the desired distinction between, on the
one hand, sentences like the ones in (9)-(10) and, on the other, sentences which are bad and
should be ruled out. This is, very roughly, the line followed both by Tanya Reinhart (in, for
instance, Reinhart(1983a,b) and Grodzinsky&Reinhart(1993)) and by Irene Heim (in
Heim(1993)). The main difference between the two is this: Reinhart distinguishes, among the
cases to which Standard Binding Theory applies, between the cases where there is binding and
cases where there is merely co-reference, and proposes that the disjointness conditions should only constrain cases involving binding, whereas cases involving co-reference should be accounted for by another sort of principle (her Co-reference Rule), which is less restrictive than principles (B) and (C) of the Standard Binding Theory. Heim, on the other hand, proposes that the disjointness conditions apply to both the cases involving binding and the cases involving co-reference; the cases involving co-reference, though, will be in addition subject to another principle (her Co-determination Rule), which can redeem some sentences that the binding principles would rule out.

I think that each of these proposals provides valuable insight into the phenomena. In this paper, though, I will not comment any further on these or other theories. I would rather like to propose here another way of looking at the phenomena we have been discussing.

3.2. Overview

An adequate theory should not only account for the contrast between the sentences which are not acceptable and those which are at least partially acceptable but it should also give an account of why the sentences in (9)-(11) have different degrees of acceptability, and specially, it should account for the difference between fully grammatical sentences like (9a), (9f) or (10b) and

\[\text{\footnotesize \textsuperscript{4}}\text{ Of course, a complete defense of the proposal I will put forward would require a careful discussion of the alternatives. I intent to present such a discussion elsewhere. We should take what we do in this paper not as an argument for the inadequacy of other theories but rather as a defense of the existence of yet another promising approach.} \]
sentences which are regarded as grammatically awkward and only partially acceptable --like, for instance, (9c) or (10d).

The proposal defended here does not consist in trying to find a set of disjointness conditions with which none of the sentences in (9)-(11) is in conflict, but rather in taking the disjointness conditions to be part of a more general account of what makes us regard some sentence (with some specific interpretation) as acceptable. This general account will explain why these sentences are acceptable even if some of them will be in conflict with the disjointness conditions that we will provide.

In the present proposal I attempt to explore the idea in Chomsky(1993) of regarding the disjointness conditions as semantic principles about how to interpret the NPs in a sentence\(^5\).

In summary the present proposal is the following: The sentences that are fully acceptable can be divided in two groups: identity sentences, and non-identity sentences. I formulate [in section 3.3] the disjointness conditions (in semantic terms) so that they do not constrain the relation of co-reference but rather the relation of presupposed co-reference\(^6\) --in this I follow Postal(1970) and Heim(1993). These disjointness conditions are then no longer in conflict with...

\(^5\) All the main ideas put forward in the following pages, though, could be incorporated, with some adjustments, into an account using a 'syntactic version' of the disjointness conditions.

\(^6\) Actually, once we take also into consideration quantificational NPs, the relation being constrained is that of presupposed co-valueness. These notions are explained in the next section (section 3.3).
fully acceptable non-identity sentences like (9f) or (10b) [section 5.1]. On the other hand, I argue [in section 5.2] that identity sentences like (the second clause of) (9a) are not constrained by the disjointness conditions, since one of the NPs is not an argument but a predicate.

One of the main contentions of this article is about what explains the (partial) acceptability of sentences like (9c) or (10d): when interpreting an utterance of a sentence like (9c) the speakers use several sources of information, the disjointness conditions being just one of them. If the other sources determine an interpretation for the NPs in a clear enough way, the sentence will be able to be understood in accordance with that interpretation even if it conflicts with the information provided by the disjointness conditions. The existence of this conflict will explain that the sentences are not regarded as completely good [sections 4 and 6].

I argue as well [in section 7] that we must distinguish two different levels in the interpretation of a sentence: one where only the lexical items and the grammatical structure of the sentence is taken into account, and a second level where contextual information is introduced so as to produce the full interpretation of the sentence.

As we just mentioned, we formulate the disjointness conditions in terms of the relation of presupposed coreference. Something is presupposed only with respect to a context. In the final section of the paper [section 8] we examine some examples that make clear the need for adequately identifying what the relevant context for the application of the disjointness conditions is.
3.3 Disjointness Conditions

In this section I propose a different formulation of the disjointness conditions. First I state the principles, then I explain the notions involved in their formulation.

I think that we should understand the disjointness conditions in the following way:

(B)' If a sentence whose LF representation is of the form ...α...β... (where β is a pronominal, and α is an NP that c-commands β in its GC) is used in a context C, it is not presupposed in C that α and β are co-valued.

(C)' If a sentence whose LF representation is of the form ...α...β... (where β is an R-expression, and α is an NP that c-commands β) is used in a context C, it is not presupposed in C that α and β are co-valued.

Let's explain what we mean by the notions of 'being presupposed', 'context' and 'being covalued' that we use in stating our disjointness conditions.

'being covalued with': NPs can have different sorts of semantic value. Corresponding to these different sorts of semantic value, there are different possible relations among NPs in a sentence.

If two NPs refer to the same individual, we will say that they are in the relation of common reference. If two NPs are in the same relation that *his* and *Every boy* are in (12) when
we understand the sentence so that every boy loves his own mother --where one NP acts like a variable bound by another NP--, we will say that the latter \textit{links} the former.

(12) Every boy loves his mother

Our characterization of the distinction between the two sorts of relation is quite uncommitted, but it suffices for the present purposes of defining the relation of 'being covalued', since, as we will presently see, the two former relations get combined in the definition of the latter.

We define 'to be covalued with' as the transitive closure of the relation 'to link, to be linked to, or to be in common reference with'. That is, \textit{NP $\alpha$ is co-valued with NP $\beta$} iff $\alpha$ belongs to any set $s$ which is such that:

(i) $\beta$ belongs to $s$ and

(ii) if $x$ belongs to $s$ then any $y$ such that $x$ links $y$ or $y$ links $x$

or $x$ is in common reference with $y$ also belongs to $s$.

\footnote{I borrow the terms "linking" and "common reference" from Higginbotham(1983,1992). My use of them is not exactly his.}

\footnote{For our present purposes, for instance, we do not need to characterize the two relations in a way which is precise enough as to decide whether in (i) (when it is interpreted as meaning that Albert loves his own brother) \textit{Albert and his} are in common reference, one links the other, or the sentence is ambiguous with respect to which of the two sort of relations holds}

(i) Albert loves his brother

\footnote{This notion of 'co-valueness' is a very close relative of the notion of 'codetermination' in Heim(1993).}

-31-
So, for instance, if in (13) we were to interpret both he and him as linked to Every boy, then he and him would be co-valued (even though the two pronouns are not themselves in a linking relation or in a common reference relation), and so such an interpretation would violate disjointness condition (B)'

(13) Every boy thinks that he loves him

'Presupposition' and 'context': The notions of 'presupposition' and 'context' that we use in stating our disjointness conditions are the same as in Stalnaker(1973,1974): linguistic communication always takes place on the basis of a background of common beliefs and assumptions, or context. Using the possible worlds framework we can identify a context in which some instance of linguistic communication takes place with a set of possible worlds: those worlds that as far as the participants in the conversation can tell could be the actual world. The context set consists of those worlds that could be the actual world according to what the participants in the conversation believe, and believe that the others believe, and believe that the others believe that they believe and so on. (If we wanted to make this characterization more precise we should take into consideration that what is relevant is not only what the participants believe but also what they pretend to believe). A proposition \( p \) is presupposed in a context \( C \) if \( p \) is true in each world in \( C \). A sentence \( s \) carries the presupposition that \( p \) if it would be infelicitous to assert \( s \) in any context where it is not presupposed that \( p \).

The individual that a referential NP picks up at a world in the context \( w \), is the individual
that the NP would refer to if \( w \) were the actual world. A sentence is *true in a world in the context* \( w \), if the sentence would be true if \( w \) were the actual world.

Let’s consider one example. Suppose I utter the sentence *He is an artist* while pointing to a man who is in front of us. Since all the participants in the conversation will believe (and believe that the others believe, etc.) that the man is in front of us and that I uttered *He* while pointing at him, it will be part of the context that the man is in front of us and that he is the one I am referring to by *He*. That is, for each world in the context \( w \) *He* will pick up that man in \( w \). Let’s suppose further that we are unsure whether the man is Jim Harris. That means that there will be some worlds in the context where the man in front of us is Jim Harris, but there will also be some worlds in the context where Jim Harris is someone other than the man in front of us. *He* will pick up Jim Harris in those worlds where Harris is the man in front of us, but will not pick up Jim Harris in those worlds where Jim Harris is someone else. This agrees with the intuitive idea that, if we do not know whether the man is or not Jim Harris, then we are unsure as well as to whether *He* refers to Jim Harris or to someone else.

---

*10* If I were to utter *He is Jim Harris*, then this sentence would express the necessary true proposition in those worlds where the man is Jim Harris and the necessary false proposition in those worlds where Jim Harris is someone else. Not knowing which worlds in the context agree with how the actual world is, we would not know whether the utterance was necessarily true or necessarily false. Still the utterance would be informative since it would indicate to anyone who accepted it, that those worlds where the man is not Jim Harris (and where the utterance expresses the necessarily false proposition) are not compatible with what we take to be the case, and so should no longer be regarded as part of the context. This picture helps clarify how it is possible that an utterance expresses a necessary proposition but it is nevertheless informative. For a more detailed explanation see Stalnaker (1979).
We are finally in the position to understand the central notion in our formulation of the disjointness conditions: Two NPs are *presupposed to be co-valued* if they are co-valued in each possible world in the context. So if $\alpha$ and $\beta$ are referential NPs then it is not presupposed that $\alpha$ and $\beta$ are co-valued if there is at least one world $w$ in the context set such that: the individual that $\alpha$ picks up in $w$ is not the same as the individual $\beta$ picks up in $w$ (i.e., if $w$ were the actual world then $\alpha$ and $\beta$ would not refer to the same individual). If a NP $\alpha$ is understood as linked by another NP $\beta$, this fact is not dependent on the context in the way that the specific value of a referential NP depends on the context, and $\alpha$ will be linked to $\beta$ in each world in the context, and so it will be presupposed that they are co-valued. So, if two NPs are not presupposed to be co-valued, then it can not be that they are interpreted so that one of them links the other.

To end section 3, let's point out that the disjointness conditions formulated in this section incorporate three kinds of ideas: First, the disjointness conditions are semantic principles that specify what the semantic significance of certain structural relations among NPs is. They are not syntactic principles that rule out certain syntactic representations. Second, the relation the disjointness conditions are about is the relation of *presupposed* co-valueness (as opposed to the relation of actually being co-valued)\(^{11}\). Third, the principles apply both to sentences containing quantificational NPs, and to sentences containing only referential NPs.

\(^{11}\) As mentioned above, I take the idea of appealing to what it is *presupposed* rather than to what is actually the case in trying to deal with the phenomena related to the disjointness conditions from Heim(1982,1993). Heim, in turn, credits Postal(1970) for this idea. The main difference between Heim's formulation of the disjointness conditions and mine is that she states her Binding Principles as purely syntactic principles that rule out certain co-indexations among NPs.
4. The Disjointness Conditions and Other Sources of Semantic Information

In this section I explain what I think accounts for the acceptability of sentences like (9b) or (10a) which are somewhat awkward but which are good enough as to pose a problem to Standard Binding Theory. The explanation in this section will make no essential use of the fact that we have formulated the disjointness conditions in terms of the relation of presupposed co-valueness. The importance of so formulating the disjointness conditions will become apparent in the next section (section 5) when we discuss those sentences that, like (9f) and (10b) are completely good.

We will introduce the idea we want to put forward in this section by means of an analogy.

First, though, a comment regarding the numbers between square brackets that will appear after some of the sentences from this point on: they are the mean grade that the sentence obtained on a 0-7 scale according to the grammaticality judgments provided by a number of speakers. We will make use of these data at several points in this paper. Even though the specific mean grade for a certain sentence is not in itself very significant (different speakers may, for instance, have used different criteria regarding how good a sentence should be in order to be given, say, a 5), the relative grade of a sentence with respect to other sentences is, I think, very significant. I think that it is very difficult to use reliable data on the relative acceptability of different

---

12 I submitted two questionnaires to 11 English speakers, asking for their linguistic judgements on a total of 36 sentences. The speakers were asked to give a grade to each sentence according to how good or acceptable they felt the sentence to be; 7 being the grade for a perfectly good sentence, and 0 the grade for a completely unacceptable one. They were asked to evaluate the sentence without thinking of any special context where the sentence might be uttered, unless such an special or uncommon context was explicitly provided in the questionnaire.
sentences without having recourse to some methodology of the kind I have used\textsuperscript{13}.

4.1. Analogy

I think it would help to explain what the present proposal is if we consider an analogy. Imagine that an English speaker is asked to determine whether the sentences in (14) below are acceptable, and that she is told that an index $m$ is meant to indicate that we are considering a reading of the sentence where the expression with the subindex $m$ refers to some individual who is a male\textsuperscript{14}

(14) a. Shem is wearing a nice dress  
b. I didn’t mean to hurt her\textsubscript{m}

The speaker would say that, given that subindexation, (14a-b) are unacceptable since these sentences can not be used with she or her referring to a man.

If we place the sentence in an appropriate context, though, the speaker might say of each of the two sentences that it is, at least, partially acceptable. For instance, if we postulate that the sentence (14a) is uttered in a situation in which a male friend has put on a woman’s dress: even if all the people involved in the conversation are well aware that the speaker is talking about a

\textsuperscript{13} For a very interesting discussion and criticism of the methodology employed in linguistics see Schutze(1995).

\textsuperscript{14} Notice that here, unlike what was the case in previous examples, the index is not part of the syntactic representation of the sentence, but just a device to indicate to the person giving linguistic judgments what is the interpretation we want her to consider.
man, the use of *she* to refer to him is acceptable. Similarly for (14b): if a rather male chauvinistic speaker wants to suggest that some man who was offended by the speaker and is crying is too sensitive, he might use (14b) where *her* will be understood by all the participants to refer to the man who is crying. In (14a-b) the use of *She* and *her* carries the information that the speaker is referring to a (human) female. This information, though, might be neutralized by some extra information provided by the context in which the sentence is uttered.

What I want to suggest is that some of the examples involving violations of Principles (B) and (C) are, in part, similar to those above involving the use of *She* and *her*: An English speaker is asked to evaluate some sentence with certain intended pattern of coreference, like, for instance, *Oscar loves him* with *Oscar* being co-referential with *him*. The sentence is declared as unacceptable since the arrangement of the NPs carries the information that they can not be assumed to corefer, that is, carries some information that contradicts what we have stipulated is the intended interpretation we are considering. If we place the sentence in a suitable context, though, for instance in (10c) [*What do you mean no one loves Oscar? Oscar loves him*], the information provided by the different semantic factors that act on that context might overcome the information provided by the arrangement of the two NPs in the last clause, and the person providing linguistic judgements will say that it is possible to understand the sentence with coreference in that particular situation, i.e., that the sentence (with the interpretation involving coreference) is at least partially acceptable.
4.2. **The Interaction of Different Semantic Factors**

We could try to account for the data we have considered in section 2 in the following way: instead of trying to provide some principles that forbid certain sort of relation (be it co-indexation [as in Standard Binding Theory], sameness of sense or *codetermination* [as in Heim(1993)], *common reference* [as in Higginbotham (1992)], or *referential dependency* [as in Evans(1980)]) when certain structural relation among NPs obtains, I think we could rather look at the disjointness conditions as describing what information certain structural relations among NPs introduce. That is, instead of looking at it from the restrictive side (what can not be the case and will be ruled out), we could look at it from the side of what it is contributed (what information is introduced by certain structural relations). This allows us to see the Disjointness Conditions as just one element in a more general picture: that of the different semantic factors that give information about the interpretation of the NPs in a sentence. Then, I content, we have a better and more natural way of explaining the grammaticality judgements of the speakers about some of the problematic sentences: When interpreting the NPs in a sentence we have, on the one hand, the information provided by the Disjointness Conditions (on the basis of the structural relations among the NPs); on the other hand, we have the information provided by other aspects of the sentence or the discourse, and by the context. These two sources of information usually concur, but sometimes they might be in conflict. If they are in conflict and if the second sort of information is strong and unequivocal enough it might overcome the information provided by the Disjointness Conditions. Then the speakers will judge that the interpretation of the sentence induced by the second source of information is possible. Nevertheless, and because of the conflict with the Disjointness Conditions, the sentence will be judged as somewhat awkward or only
partially acceptable.

Sentence (15) exemplifies this point. Consider (15) (which is the same as (10a)):

(15) Mary, Betty and John have one thing in common: Mary admires John, Betty admires him, and John himself also admires him. [3.8]

The structural arrangement of *John himself* and *him* in the last clause carries the information that we can not understand *John himself* and *him* so that it is presupposed that they are the same individual; on the other hand, other factors force that we realize that the individual the two NPs will refer to is the same. The final result is that we can understand the sentence with the coreferential interpretation of *John himself* and *him*. Because of the conflict between the information that different aspects of the sentence convey, though, the sentence has a somewhat odd character (its mean grade is [3.8]). The factors that force that both *John himself* and *him* are taken to refer to John are: *John himself* refers naturally to John --the same person named *John* that has been referred to in the previous clauses--; this is reinforced by the fact that we are expecting to be told what Mary, Betty and John have in common, and in the previous two clauses the subject NP refers to, respectively, Mary and Betty, and so we expect the subject NP in the last clause to refer to John; in addition the use of *himself* also forces *John himself* to refer to the same male named *John* that the previous clauses refer to, since the use of *himself* carries the presupposition that the person in question has already been mentioned. The use of *also* in the last clause introduces the presupposition that a property that is being attributed in the last clause has
already been considered (the last clause can be seen as involving the attribution of either one of
two properties: that of 'admiring him', and that of 'being admired by John'); the repetition in the
previous clauses of the attribution of the property of admiring John, makes that property
completely salient in order to be the property that also carries the presupposition about (as opposed to the property of being admired by John); now, if also in the last clause introduces the
presupposition that the last clause is attributing the property of admiring John, that means that
him in the VP of the last clause should refer to John. Finally, the fact that it has been announced
at the beginning that we would be told what Mary, Betty and John have in common, and that we
have already been told that what Mary and Betty both do is to admire John, makes us to expect
that in the last clause we will be told two things: first (as we already mentioned above) what
John does, and, second, that what he does is the same that Mary and Betty do: to admire John.
That means, again, that both John and him will have to refer to John. All these factors force us
to interpret the sentence so that we realize that both John himself and him refer to the same
individual. That is, all these factors together counteract the information conveyed by the
disjointness conditions on the basis of the structural relation between the two NPs.

Notice that if the sentence were (16)

(16) Susan admires John, and he admires him

the sentence would be quite acceptable but the last clause would not be understood so that the
two NPs refer both to John but rather it would probably be taken to mean that John admires
some man or boy who, it would be assumed, the speaker is treating as having already been introduced in the conversation. If asked whether the last part of (16) can mean that John admires himself, the subject offering her linguistic judgements would have to say that, if nothing else is added, it can not. If forced to give a grade to the sentence 'when there is coreference' it would be a very low grade (16) --with the addition of italics to indicate intended coreference-- obtained a mean grade below 1).

When a speaker is asked to evaluate a sentence like the last clause of (16) (with no particular context) coreference is deemed unacceptable (since the only factor that gives information about the relation between the two NPs is the structural arrangement of the two NPs, and the information that this arrangement gives is that they can not be taken to corefer); but if a discourse or context is added which provides additional information about who the two NPs refer to, the speaker might judge that coreference is possible, and so that a co-referential reading of the sentence is, at least, marginally acceptable. Something similar was the case with respect to she and the possibility of referring to a male.

Notice that from the account of the (partial) acceptability of (15) that the have provided, it trivially follows an explanation of why if we modify (15) so as to suppress some of the features that we have said help the sentence to be regarded as good the sentence becomes less good. For instance if we omit himself or also as in (17a) the sentence becomes less good; it
becomes even worse if, in addition, we do not mention one of the women as in (17b)\textsuperscript{15}. Also, if we omit to announce that we are going to tell what it is that the three people have in common, as in (17c), the sentence is certainly worse than (15).

(17) a. Mary, Betty and John have one thing in common: Mary admires John, Betty admires him, and John admires him.

b. Mary and John have one thing in common: Mary admires John, and John admires him.

c. Mary admires John, Betty admires him, and John himself also admires him. [2.2]

We will further develop the idea introduced in this section in section 6, where we will comment on some of the other sentences in (9)-(11) that are regarded as acceptable for the same reasons as (15) is.

5. Identity and the Disjointness Conditions

In this section we will focus on those sentences analogous to those in (9)-(10) that are completely good ((9a), (9f) and (10b)). We will see that their interpretation does not involve any conflict with (our formulation of) the disjointness conditions.

\textsuperscript{15} Unfortunately, the linguistic judgements questionnaire I use did not include the totality of the sentences we will use in our discussion. In particular, it did not include sentences (17a) and (17b). I will have to hope that the intuitions of the reader with respect to these two sentences coincide with my own --which are the ones expressed in the text.
5.1. Non-presupposed identity

Under our formulation of the disjointness conditions it is easy to explain why the last part of (10b) and of (9f) are completely acceptable sentences and, unlike some of the other sentences in (9), (10) and (11), they do not have any flavour of oddness or of being only partially good. Consider, for instance, (10b) (repeated here as (18))

(18) At a Halloween party someone says: I think that the man with the devil costume is Joe. It is suspicious that he knows him so well.

The reason why it is completely good even if he and him in the last clause actually refer both to Joe is this: when uttering the last part of (18) it is still an open question whether the man in the devil costume (who is the one he refers to) is or is not Joe (who is the one him refers to). Whether they are the same or not is precisely what is being discussed. Putting it in terms of possible worlds: in some worlds in the context, Joe is the man that is wearing the devil costume at the Halloween party, but in some other worlds in the context someone other than Joe is the one who is wearing the devil costume. That means that we interpret he and him so that they refer to the same individual in those worlds in the context where Joe is the one wearing the devil costume, but they refer to different individuals in those worlds where someone else is wearing the costume. So, we interpret the two NPs so that it is not presupposed that they refer to the same individual (and so it is not presupposed that they are co-valued). There is no problem, then, in interpreting the last part of (18) in accordance to (B)'.

-43-
We could not interpret "He knows him so well" in (18) with He referring to whoever is the man in the devil costume, and him referring to Joe if it were not an open question whether the two individuals are the same or not. We can see this in (19), where the previous discourse has been modified so as to try to make clear that when the utterance of the last clause takes place it is presupposed that Joe is the man with the devil costume.

(19) A: Do you have the list of who is each person in the party?
B: I certainly do.
A: Could you tell me who is the man with the devil costume?
B: The man with the devil costume is Joe
A: Oh, I see. It is suspicious that he knows him so well.

We can not understand the last clause so that he and him both refer to that individual that we have already established that is both Joe and the man with the devil costume.\(^\text{16}\)

\(^{16}\) We can make the sentence good, or at least much better, by changing the tense of the discourse to past, and by replacing "It’s suspicious that" in the last clause by an expression such as "This is why", "This explains that" or "No wonder then that". I do not think this poses a problem for the claim that what makes the last part of (19) good is that we are not presupposing that the individuals that he and him refer to are the same. The use of past tense and of expressions like "This is the reason why..." facilitates interpreting the sentence with respect to the context as it was before the previous sentence had been uttered. We can also see this in a text like (i):

(i) A: I wonder why Tom did not come to the party with his wife.
B: Tom has never been married.
A: This explains why he did not come with his wife.

The use of his wife in the last clause requires that it is not presupposed that Tom is not married. This, though, is exactly the information that has been introduced in the context by B.
5.2 Identity Sentences

Regarding sentences with a so called equative use of the copula like for instance the last part of (20) (which is the same as (9a))

(20) Who is that woman over there? She is professor Rigau

one might think that the explanation for why they are completely good sentences is the same that we just gave for (18). However this is not really the case.

5.2.1. A Problem. It is true that usually when we utter a sentence such as the last clause of (20) we are not presupposing that the two NPs refer to the same individual. Nevertheless it still might be perfectly fine to utter an identity sentence in a situation where according to what the participants assume, the two NPs will refer to same individual. For instance, if Diana, Pilar and Colin are in a room and they all know each other (and know that they know each other, etc), Diana’s utterance of (21) is still good

(21) He is Colin

Nevertheless, the use of "this explains why" makes easier to understand that what follows does not take for granted the information that this (in "this explains why") refers to.

More could be said about how expressions like "this is the reason why..." affect what the participants will be able to understand is the context they should use to evaluate a particular utterance. Examining this any further, though, is beyond the scope of this paper.

-45-
It might be a silly and pointless utterance, but it is still perfectly good from a grammatical point of view.

5.2.2. *Heim’s Reply.* One reply given to this objection by several authors (in Heim(1988,1993), Lasnik(1990), and in some way also in Higginbotham(1992)) consists in claiming that by the very act of uttering a sentence that asserts certain identity the speaker indicates that we are not presupposing that identity (and so that the hearers should assume that the identity was not presupposed, even if they had thought it was\(^{17}\)). If the context is assumed to have been such that it was not presupposed that the two individuals were the same, then the identity sentence is not in conflict with our disjointness conditions.

The reason why the speaker indicates we are not presupposing the identity when uttering the identity sentence is that it is a general principle governing conversation that what it is asserted can not be presupposed. This principle follows from Grice’s Cooperative Principle, and, in particular, from the maxim of Relation\(^{18}\).

5.2.3 *Objection-1 to Heim’s Reply.* I do not think we should be satisfied with this explanation of why identity statements are perfectly good sentences even when the identity is already presupposed. One way of realizing that this explanation can not be satisfactory is by noticing the

\(^{17}\) This process of reacting to an utterance of the speaker by regarding the context as having been different from what it actually was is what David Lewis calls *accommodation.* See Lewis(1979).

\(^{18}\) See essay 2 in Grice(1989).
following: The principle that what is said can not be presupposed is just a principle about rational interchange of information. As with any other principle derived from the Cooperative Principle, it is possible, at the price of acting silly, not to act in accordance with the principle. I can say to you "This man is Chomsky", and you might reply "Yes certainly, he is Chomsky", and I might go on to say "He is Chomsky. He is Chomsky". My utterance would be silly, but completely acceptable from a grammatical point of view. The reason why we feel that my utterance would be silly is that it would violate the maxim of Relation ('Be relevant'). So, if in the situation described we would realize that the maxim of Relation is not operating, we should not expect the principle "what is said can not be presupposed" (that gets its justification from that maxim) to apply either. Nevertheless, the sentence is good. So, it is not the operation of the principle "what is said can not be presupposed" that explains why in the situation described the last utterance of "He is Chomsky" is still completely correct.

5.2.4. Objection-2 to Heim's Reply. The reasoning just given shows that it can not be that the explanation for why identity sentences are correct in contexts were the identity is already presupposed makes essential use of the maxim of Relation, because there are contexts where the maxim of Relation is violated but where an identity sentence stating some identity that is already presupposed is still acceptable. Now, a more direct way of objecting to what we have termed Heim's Reply is to indicate where exactly there is a problem in the argumentation given as part of the Reply. In order to do so, it will help to make more explicit what the argument that appeals to the principle 'what is said can not be presupposed' and that I am objecting to is.
The argument can be formulated in the following way: Let NP1 and NP2 be two referential noun phrases, and C a context where 'NP1 is NP2' is uttered and where the referents of NP1 and of NP2 are presupposed to be the same; let val(X,w) denote the individual that the referential expression X picks up in the world w, then, by assumption, we have (i) and (ii)

(i) 'NP1 is NP2' is uttered in C
(ii) NP1, NP2 and C are such that: \( \forall w \in C (\text{val}(\text{NP1},w) = \text{val}(\text{NP2},w)) \)

Given (ii) we have that what 'NP1 is NP2' says is true in each world in C, and so that (iii) holds\(^{19}\)

(iii) What 'NP1 is NP2' (when uttered in C) says is already presupposed in C.

By Grice's Cooperation Principle and, in particular, by the maxim of Relation, we have (iv)

(iv) When the speaker utters 'NP1 is NP2' in C, she is not saying something that is already presupposed in C.

So, there is a conflict between (i), (iii) and (iv). It is a general fact that conflicts involving the Cooperation Principle can be solved by accommodating (i.e., by assuming that the context was different from what it actually was). In the present situation the conflict is resolved by assuming

\(^{19}\) Recall that a context C presupposes that p if for each \( w \in C \), p is true in w.
(v) The context where 'NP1 is NP2' was uttered was not C, but rather a context C' which is such that $\exists w \in C'(\text{val}(NP1,w) \neq \text{val}(NP2,w))$

Finally, given (v) we have (vi)

(vi) The context where 'NP1 is NP2' is taken to have been uttered is such that there is no conflict between the disjointness conditions and the interpretation of the utterance in that context.

Now that we have the argument laid out in detail we can see that there is a problem in (iii). (iii) assumes that the sentence 'NP1 is NP2' has an interpretation before the conclusion (vi) which establishes that the sentence is not in a conflict with the disjointness conditions has been reached. Premise (iii) could be analyzed into two sub-premises (for concreteness we assume that what a sentence says is a proposition, that a proposition is a set a possible worlds, and that identity statements express the diagonal proposition in the sense of Stalnaker(1979) --these assumptions are not at all essential to the point we want to make, though$^{20}$):

$^{20}$ We could take (iii)a and (iii)b simply to be

(iii)a. 'NP1 is NP2' when uttered in C says that p.
(iii)b. C presupposes that p.
(iii)a. An utterance of "NP1 is NP2" in C expresses the proposition:

\{w \in C: \text{val}(NP1, w) = \text{val}(NP2, w)\}

(iii)b. C presupposes the proposition \{w \in C: \text{val}(NP1, w) = \text{val}(NP2, w)\} (since from (ii) it follows that C = \{w \in C: \text{val}(NP1, w) = \text{val}(NP2, w)\}, and so, in particular, C is included in \{w \in C: \text{val}(NP1, w) = \text{val}(NP2, w)\} )

(iii)a is completely unjustified, unless we regard the problem we are trying to solve as already solved --that is, unless we assume that there is no problem in establishing that, given certain context, the identity sentence has certain interpretation even though the disjointness conditions prescribe that the sentence does not have that interpretation given that context. (Notice, incidently, that the disjointness conditions are not taken at all into consideration in the argument (i)-(v) that establishes conclusion (vi)).

In particular, under the assumptions accepted by Heim(1982, 1993) (where the disjointness conditions are syntactic constrains ruling out certain syntactic structures) (iii) is certainly not true. Any syntactic representation of "NP1 is NP2" that would give rise in C to the interpretation that (iii) attributes to the sentence would already be ruled out at the syntactic level by the Principles (B) or (C) of the Binding Theory\(^{21}\), and so would not express any proposition.

\(^{21}\) Heim(1993) focuses her discussion only on phenomena related to principle (B). What she says, though, is easily and naturally extendable to phenomena related to principle (C) as well.
5.2.5. Another Reply. Maybe behind what I called Heim's reply lurks another argument that does not appeal to the principle 'what is asserted can not be presupposed'. It is the following: Suppose that a context C is such that a use of the pronoun he would naturally pick up Colin in each world in C, and so would a use of the NP Colin. If someone utters He is Colin in C, what would be the natural way of interpreting the two NP conflicts with what the disjointness conditions prescribe. The hearers, assuming that the speaker is rational and is trying to abide by the Cooperative Principle, will try to find a way of escaping the conclusion that the speaker's utterance is not interpretable. There is one way of doing so: to accommodate, that is, to assume that the context was not C but rather C', where C' is such that for some world in C' the two NPs pick up different individuals. So the hearers will accommodate, and if accommodation takes place then there is no problem in interpreting the sentence in accordance with the disjointness conditions. So, the fact that accommodation will take place explains that an utterance of He is Colin is grammatical even if uttered in a context where the identity was already presupposed.

I think, though, that this argumentation is not satisfactory. If the reasoning in the previous paragraph were correct then accommodation should also be an option when trying to interpret other sentences that also involve a conflict with the disjointness conditions. Consider the contrast between (22) and (23)

(22) This woman is Pilar. She is Pilar.
(23) This woman is Pilar. She knows Pilar.
(23) is quite bad (if \textit{She} is taken to refer to Pilar), whereas (22) is fully acceptable. If accommodation is possible when interpreting the second clause of (22), then it should also be possible when interpreting the second clause of (23); and if the fact that accommodation is possible is what explains that (22) is good then (23) should be predicted to be also perfectly good; but (23) is not good.

Similarly, we can not explain \textit{my} identity sentences of the same form as (21) \textit{[He is Colin]} are completely good by appealing to an explanation like the one we gave in section 4 to account for the acceptability of sentences like (10a) or (10c)\textsuperscript{22}. Sentences like (10a) or (10c) are only partially acceptable. As pointed out in section 4.2, it is precisely the fact that the interpretation of these sentences is in conflict with what the disjointness conditions prescribe that explains why they are only partially acceptable. Identity sentences even when uttered in contexts as the one described for (21) are perfectly good, though. So, it can not be that the interpretation of identity sentences involves a resolution of a conflict similar to the one involved in the interpretation of (10a) or (10c). So, the explanation of the complete acceptability of identity sentences can not be along the same lines as the one we gave for (10a) or (10c).

\textsuperscript{22} Recall that the explanation was that the interpretation of the sentence involves a conflict between the disjointness conditions and other kinds of information; In sentences like (10a) and (10c) the information opposing the disjointness conditions is strong enough as to prevail and make the sentence interpretable.
5.2.6. A feasible solution. I think that a promising way of trying to account for the complete grammaticality of identity sentences is in terms of the special syntactic character of sentences containing the verb to be. Several works have pointed out and tried to explain the special syntactic properties of the so called copular sentences\textsuperscript{23}. In particular Heggie(1988) argues that the predicate of copular sentences is not the verb to be, but rather one of the NPs\textsuperscript{24}. That means that there is only one argument in identity sentences: the NP that is not the predicate\textsuperscript{25}. If we assume that our disjointness conditions apply only to arguments, then (B)' and (C)' do not provide any information about how copular sentences like (21) should be interpreted. The interpretation of identity sentences does not, then, involve any conflict with the disjointness conditions.

Discussing the different arguments that show the special character of copular sentences


\textsuperscript{24} According to Heggie this claim requires some qualifications.

\textsuperscript{25} Argument and predicate are technical notions in linguistic theory. They are usually characterized in terms of the so called Theta-theory. One alternative way of characterizing the notions of argument and predicate is in terms of the type of semantic values they can have. We will not commit ourselves here to any particular characterization of the two notions. Among many other alternative ways, one of the simplest characterizations would be the following:

We define a predicate of degree-n inductively as follows: A predicate of degree-1 is an expression whose semantic value is a function from the set of individuals to the set of truth values; a predicate of degree-n+1 is an expression whose semantic value is a function from the set of individuals to the set of predicates of degree-n. X is a predicate if, for some n, X is a predicate of degree-n. X is an argument, if X can be functionally combined with a predicate of degree-1 to yield a truth value. (Notice that under this definition not only expressions whose semantic value is an individual can be arguments; if the semantic value of a quantificational expression is a function from predicates of degree-1 to truth values, then quantificational expressions are also arguments).
and that try to prove that the predicate of copular sentences is one of the NPs, as well as examining the possible evidence against this view is by itself a lengthy topic. The reader is referred to the discussion in Heggie(1988) and the other works mentioned in footnote 23. Here we will restrict ourselves to mentioning one kind of data that gives plausibility to the view, without entering in any further details.

In Catalan the clitic *el* (*l’*) corresponds to an argument position; the clitic *ho* can correspond to predicates but can not correspond to any personal NP in an argument position. This is illustrated in (24): in (24b) the clitic *el* stands for the argument *el Joan* in (24a), while it is not possible for *ho* to stand for that argument.26

(24)  a. Aquell home estima *el Joan*
      that man loves (the) John

      b. Aquell home *l’/*ho estima
      that man him/it loves

In (25b) the clitic *ho* stands for the predicate *molt feliç* ’very happy’ in (25a), whereas *el* can not stand for that predicate

26 Heggie(1987) and Longobardi(1985) describe similar facts to (24)-(26) for French and Italian, respectively.
(25)  a.  El Joan és molt feliç
       (the) John is very happy

   b.  El Joan *l'/ho és
       (the) John him/it is

In contrast with (24a-b), the argument clitic el in (26b) can not be made to stand for the NP el Joan in the identity sentence (26a), but the predicate clitic ho can:

(26)  a.  Aquest home és el Joan
       this man is (the) John

   b.  Aquest home *l'/ho és
       this man him/it is

The data (24)-(26) suggest that the NP el Joan in (26a) is not an argument but a predicate. If the disjointness conditions apply only to argument NPs, then the disjointness conditions are irrelevant for the interpretation of (26a). In particular, there would be no conflict with the disjointness conditions if we interpreted the two NPs in (26a) in such a way that they determine the same individual in each world in the context.\textsuperscript{27}

\textsuperscript{27} I use the word "determine" in an ambiguous way. The two NPs will 'determine' the individual in a different way, since, presumably, they will have a different type of semantic value. If the semantic value of the argument NP Aquest home 'this man' is an individual, then the semantic value of the predicate NP el Joan 'John' might, for instance, be a set of individuals --actually, a \textit{singleton} set, which also \textit{determines} an individual.
6. Some comments

6.1. The Disjointness Conditions and the Cooperative Principle

That our Disjointness Conditions (B)' and (C)' be concerned with what it is presupposed to be the case, rather than with what is actually the case, is what allows our proposal to deal with examples like (18). It could seem, though, that this very same feature gives rise to wrong predictions regarding the use of sentences like (27) in certain contexts.

(27) He loves him

Usually when someone utters (27) the context will contain information about who each NP refers to (say, we are talking about John’s feelings for Paul, or we are talking about who loves Clinton and at that point Henry comes into the room, etc). If this contextual information implies that *He* and *him* refer to different individuals (and so that, for each world in the context, the individual that *He* picks up is distinct from the individual that *him* picks up), then we will interpret the sentence with the two NPs referring to two different individuals. This is in agreement with what the Disjointness Conditions prescribe (the Disjointness Conditions prescribe that it not be presupposed that the NPs refer to the same individual, and the contextual information agrees with that since this information will determine that the individuals that the two NPs refer to are presupposed not to be the same28).

28 Compare: (i) Not to presuppose that p; (ii) to presuppose that not-p. (i) does not imply (ii), though (ii) implies (i) (assuming consistency), so to satisfy (ii) is one way of satisfying (i).
If we hear someone utter (27) and we have to interpret it without being able to use any previous context (say we have just joined an ongoing conversation), though, we will also think that the individuals referred to by *He* and *him* are distinct, not just that it is not presupposed that they are the same. This might seem to be a problem for the formulation of the disjointness conditions that I am defending for the following reason:

If (27) is uttered in a context that contains no information about who the referent of the two NPs is, then it seems that the hearers, in order to interpret the NPs, will have to rely solely on the lexical information and the information provided by the disjointness conditions. The lexical information is just that the individuals the NPs refer to are male and are not the speaker or the hearers though they are somewhat salient at that point in the conversation. The information provided by the disjointness conditions is that either (i) it is an open question whether the two NPs refer to the same individual (i.e., in some worlds in the context they pick up the same individual, but in some other worlds they do not), or (ii) the two NPs refer to two different individuals (i.e., in each world in the context they pick up different individuals). The disjointness conditions do not determine which of (i) or (ii) is the case.

So, using solely the disjointness conditions (and the lexical information) we could not conclude (ii), but only that either (i) or (ii). But, as pointed out above, (ii) is what we do conclude when we hear (27) in a context that does not provide any information about the reference of the NPs. This seems to suggest that there is a problem for my formulation of the disjointness conditions and that they should be amended. If what we conclude when we hear (27)
in a context that does not provide information on the reference of the NPs is that they refer to different individuals, and it seems that all the information we use in such a context comes from the disjointness conditions (and lexical semantics), then it seems that what the disjointness conditions should prescribe is that (when the relevant syntactic relation obtains) the interpretation of the two NPs must be presupposed to be distinct, not just not presupposed to be the same. (Adopting this modification would, of course, be problematic in other respects since, for instance, we would no longer be able to account for sentences like (18)).

I think, though, that the suggested objection above overlooks one kind of information that will be available to the hearers when (27) is uttered even if the context does not include any information about who the NPs refer to: that the speaker is conforming to Grice’s Cooperative Principle. Barring information to the contrary, we will always assume that the Cooperative Principle and in particular, the Gricean maxims, are respected. And we will rely on this assumption when trying to interpret what the speaker says.

If we hear (27) in a context that does not contain information about who the referent of the two pronouns is, but that, otherwise, is a non-special context, we will assume that the speaker knows who she is talking about. So she knows who the person that she is referring to by He is, and who the person that she is referring to by him is, and so she knows whether they are the same or not. If she knew she was talking about one single person, then she would not talk as if the question whether there is one person or two was open, since doing so would go against the maxim of Quantity (‘be as informative as is required’). So, since she is treating the question as
open, we can conclude that she is not talking about one single person but two.

Of course there can be special contexts (as, for instance, in (18)) in which it is clear that the speaker is unsure about the identity of the individuals she is talking about, or where she has some good reason to act as if she is unsure. This is not, though, what is usually the case. Barring information to the contrary, we will assume that the conversation does not take place in any of these special contexts, and so we will assume that the speaker knows who she is talking about, and that she is open about it. So, assuming disjointness condition (B)' and the Gricean maxims, it follows, as desired, that in non-special contexts the hearers will assume that when the speaker utters (27) she is talking about two individuals.

Notice that in a special context where it is reasonable to believe that Gricean maxims do not apply, like, for instance, when an oracle says something, or in the statement of some puzzle in, say, the Sunday edition of a newspaper, we would not conclude from the use of (27) that the individual referred to by He is not the individual referred to by him, but rather only that it might or it might not be the same.

6.2. Sources of information for the interpretation of NPs

As we have pointed out, there are several sources of information that are used in interpreting the NPs in a sentence, besides the information provided by the disjointness conditions on the basis of the structural relation among the NPs. In this subsection we will examine some of the ways
of generating the information that, as we have seen, in sentences like the ones in (9)-(11) can conflict with and prevail over the information provided by the disjointness conditions.

6.2.1. Presuppositions. The most common way of generating such information is by making it clear that the clause in which certain NP X appears expresses some property which is the same as certain property that has been introduced before. The property involves certain individual or certain pattern of linking. So X must refer to that individual or be subject to that particular pattern of linking. One way of having this effect is by repeating several times the attribution of a property, so as to create a pattern that will make the listener expect that the next attribution of a property will fit the same pattern (this is illustrated, for instance, in (9c)). Another way of having this effect is by having mentioned the application of the property as a general case and then making clear that we are considering a particular application of that general case (this is illustrated in (9e) and (11a)). One specially good way of having this effect, though, is by using some device which introduces a presupposition. We have already commented (with respect to (9c)) on the effect that the presence of also can have. The word even (that appears in (9b)) has a similar effect to that of also. The sentences in (9)-(11) illustrate, though, other ways of introducing presuppositions besides including some specific word. Consider, for instance, (28)

(28) What do you mean no one loves Oscar? He loves him. [2.5]

He clearly refers to Oscar, which is the only salient individual when the pronoun is uttered. In a more common discourse him would be taken as referring to some other individual different
from Oscar (because of disjointness condition (B), and the Cooperation Principle). But a sequence of the form What do you mean \( \alpha \)? \( \beta \) carries the presupposition that \( \beta \) implies no-\( \alpha \). In the case of (28) this means that \textit{He loves him} implies that it is not the case that no one loves Oscar. The most simple and likely way for \textit{He loves him} to have that implication is if \textit{He loves him} itself is an attribution of the property of loving Oscar. If this is so, then \textit{him} must refer to Oscar, disjointness condition (B)' notwithstanding.

Another way of introducing presuppositions is by stressing some word. Whatever exactly the presupposition induced by stressing an NP is, it includes that the property being attributed to the individual determined by the NP, has already been considered. So in (29)

(29) What do you mean John loves no one? He loves JOHN. [5.7]

the stress on \textit{John} carries the presupposition that the property of being loved by the subject of the last clause has already being considered. There are two properties that have been considered in the first clause: the property of being loved by John and the property of loving no one. The latter, though, can not be the one that the stressing of \textit{John} carries the presupposition about since in the last clause --even before determining who the NPs refer to-- it is clear that we are not attributing the property of loving no one. So the presupposition brought about by stressing \textit{John}, is that, in the last clause, we are attributing the property of being loved by John; that means that \textit{He} must refer to John, disjointness condition (C)' notwithstanding. There is also, of course, the presupposition brought about by the structure What do you mean \( \alpha \)? \( \beta \) which adds its effect to
the stressing of *John*.

### 6.2.2. Demonstration

Another way of generating information that can override the information provided by the disjointness conditions is by using an NP demonstratively. (30a-b) illustrate this

(30)  

a. He refuted HIM [points to the person]  [3.1]

b. A: John saw Peter

B: No, John saw HIM [points to John]  [5.3]

In, for instance, (30a) *He* refers to whoever is most salient individual at that point in the conversation; because of disjointness condition (B)', without demonstration *him* would be assumed not to refer to the same individual; the demonstration, though, forces it to refer to that same individual.

Using an NP demonstratively is a very clear way of indicating what the NP refers to; this is why the intended reference can be communicated even if the disjointness conditions are providing opposing information. Notice that most of the sentences considered in the previous subsection involved several devices that together were able to quite successfully override (B)' or (C)'. (30a) shows that pointing is effective enough as to have that effect on its own. Still the combination of demonstration with other devices, as in (30b), makes the overruling of (B)' clearer and the sentence better.
6.2.3 Pronominals versus Referential Expressions. A pronoun is very strongly dependent on the context to determine what it refers to. This is not so for proper names and definite descriptions. They are to an extend dependent on the context (there are many people named John and many presidents), but not as much as pronouns are. Proper names and definite descriptions are able, to a good extend, to determine their referent 'on their own'.

It is usually said that 'Principle (C) violations' are less strong than 'Principle (B) violations'. Under my proposal we can explain what motivates this claim without having to accept the queer idea that there is a different degree of prescriptiveness associated to each of the disjointness conditions.

Notice that the claim does not seem to be true of sentences violating (C) that include a pronoun but do not contain any indication at all about who the referent of the pronoun is. So, for instance, I think that without any particular previous context, it is as hard and unlikely to understand that the NPs in (31b-c) (principle (C) configurations) refer to the same individual as it is in (31a) (a principle (B) configuration)

(31) a. He admires him  
b. He thinks that Paul is crazy  
c. He admires Paul

29 As we see in the next paragraph, the proponent of a different force associated with each disjointness condition would also have the problem of explaining why some kinds of sentences violating (C) are as bad as their analogues which violate (B).
I think that what motivates the claim that 'violations of (C) are not as bad as violations of (B)' are contrasts like (32a-b)

(32) a. I talked to Paul today. He admires him
b. I talked to Paul today. He admires Paul

It seems that coreference in the last clause of (32b), even if it is still quite bad, is not as bad as in the last clause of (32a). I think we can explain why in the following way: Both in (32a) and (32b) the pronoun He would naturally tend to pick up Paul as its referent. In (32a), nothing indicates what the referent of him ought to be; disjointness condition (B)' indicates that it can NOT be assumed to the same individual He refers to; so, because of the effect of (B)', him can not be understood as also referring to Paul. In the second clause of (32b), unlike what is the case with respect to him in (32a), Paul would naturally tend by itself to refer to the Paul that is mentioned in the first clause; because of the effect of (C)', though, the most likely way of taking an utterance of (32b) (without any other relevant context), would probably be as Paul in the second clause referring to a person named Paul distinct from the Paul mentioned in the first sentence. In any case, even if the presence of a proper name, which by itself gives clear indication of what its most likely referent is, is not sufficient to neutralize and prevail over the effect of (C)', it is enough to diminish it. This makes it easier to understand the last part of (32b) as involving coreference than to understand the last part of (32a) as involving coreference; this, in turn, is what makes us feel that a coreferential interpretation in (32b) is not as bad as in (32a).
6.3. *Corroborating Data*

The grade that was given to (the coreference reading of) the sentences in (33) is exactly what should be expected if the proposal under consideration and the observations in the previous sections are correct\(^{30}\)

(33)  

a. What do you mean no one loves Oscar? He loves Oscar [2.9]  
b. What do you mean no one loves Oscar? HE loves Oscar [4.2]  
c. What do you mean no one loves Oscar? He loves him [2.5]  
d. What do you mean no one loves Oscar? HE loves him [3.6]

Sentence (a) is better than (c), and (b) better than (d), in accordance to the fact that proper names can determine its referent with much less dependence from the context than a pronoun (and so are less affected by the opposing effect of (B)' or (C)'); (b) is better than (a) and (d) better than (c), as should be expected if stress introduces the presupposition that the property attributed to the subject of the second clause is a property that has already been considered in the previous clause, and if this presupposition helps to make clearer what the referent of the last NP is, and this, in turn, makes a difference on how acceptable the sentence is. Finally, the sentences in (33) seem to suggest that those devices that can make coreference in a sentence more or less acceptable, add to each other: (d) is better than (c), and (b) is even better than (d). ((d) includes

\(^{30}\) I do not pretend that the results of my linguistic questionnaire provide a completely reliable measure of the acceptability of the different sentences. More tests would be needed to support the results I obtained. Still, I think it worth mentioning how well the grades obtained by the sentences in (33) fit with the proposal I am defending.

-65-
the use of the presupposition generating structure \textit{What do you mean }\alpha? \beta \textit{ and the use of stress, whereas (c) does not contain stress; (b) in addition to the devices in (d), also includes the use of a proper name).

7. \textbf{Two-stages in the interpretation of an utterance}

The proposal I present in this paper is not yet complete. We have seen that the basis for explaining the acceptability of sentences like (9b-e) is the fact that the interpretation of these sentences involves an interaction between, on the one hand, the information provided by the disjointness conditions and, on the other hand, other information provided by other aspects of the sentence, or by the context. We will see in this section that it is necessary to be more specific about what is involved in this interaction in order to satisfactorily account for the data. In particular we will see that we need to take into consideration the existence of two stages in the interpretation of a sentence. That something must be missing in the proposal as it stands is shown by (34)

(34) I refuted me [0]

It is completely clear that the pronouns \textit{I} and \textit{me} in (34) should refer to the same person, the speaker. So it would seem that, according to the proposal I am defending, the effect of disjointness condition (B)' in (34) should be overridden by the opposing and unequivocal information about the reference of the pronouns that the very pronouns provide. And so, (34)
should be, at least, partially acceptable. (34), though, is completely bad.

It is clear that the process of interpretation of a word, sentence or piece of discourse involves different aspects. For example, part of the interpretation of an utterance of the word *she* will consist in recognizing, for instance, that the word has been used referentially and to interpret the word as an expression that will potentially refer to some human female that is in some way salient at the moment of the utterance. This is part of the interpretation of the utterance but it is not all there is to interpreting the utterance. Another aspect of the interpretation is to determine which specific individual the word refers to (for instance, to determine that it is Delia that the speaker refers to by his use of *she*).

In order to satisfactorily account for the phenomena related to the disjointness conditions we must take into consideration the existence in the process of semantic interpretation of the two stages suggested in the previous paragraph: in the first stage only structural, lexical and general semantic information is used, and a partial interpretation is produced. For instance, a partial interpretation of a referential use of a pronoun like *he* would identify it as a referential expression that picks up a male human who is in some way salient. The partial interpretation for a sentence like *He loves him* would be the pseudo-proposition that a not-yet determined male

---

31 If we take the full interpretation of a pronoun to be a two dimension individual concept, that is, a function $f$ that to each world $w$ in the context assigns a function $g$ that assigns an individual to each possible world (in the case of a rigid designator $g$ is constant, though $f$ might not be), then a partial interpretation of a pronoun is a function $h$ whose arguments are sets of possible worlds (contexts) and whose values are individual concepts ($h$ will only be defined for some contexts: those where an utterance of the pronoun takes place, and where the language is the one that the interpreter is considering).
human loves a not-yet determined male human who it is not assumed to be the same as the first one. This partial interpretation is produced on the sole basis of the lexical information of the words and the semantic rules—including the disjointness conditions (which depend, in turn, on the structural relationship between the NPs).

In the second stage the information from other parts of the discourse and from the context is brought in and the complete interpretation is produced. For instance, in the second stage of the interpretation of an utterance of he, the contextual information that, say, John is a salient human male and the individual that is most relevant for the topic that is currently discussed is used in order to determine, on the basis of the partial interpretation of he produced in the first stage, that the speaker referred to John. In the second stage of the interpretation of He loves him, discourse and contextual information (like for instance that the topic under discussion is James’ feelings for John) is applied to the first stage’s partial interpretation in order to produce the final interpretation of the sentence (for instance that James loves John).

---

32 If the full interpretation of a sentence in a context C is a propositional concept, i.e., a function f that to each world in C assigns a function g that assigns a truth value to each possible world, then we can postulate that the partial interpretation of a sentence like He saw him is a function h that takes as an argument a set of possible worlds (a context) and that yields as a value a propositional concept. h will only be defined for some contexts: those that are such the utterance of the sentence has taken place, where the language is the one the interpreter is considering, and which are such that: the interpretation of the two NPs with respect to that context (i.e. the individual concept that, in accordance to the partial interpretation of each of the NPs, is the value of that NP with respect to that context) is such that: the diagonal of the two individual concepts is not the same (this last condition is what incorporates the idea that it is part of the partial interpretation of the sentence that we do not presuppose that the reference of the two NPs is the same).
It might be that the contextual information that is used in the second stage is in conflict with some aspect of the partial interpretation produced in the first stage. The observations about how the conflict might get resolved made in the previous sections apply here: the existence of the conflict *might* make the sentence to be regarded as bad, but also, if the opposing contextual and discourse information brought in at the second stage is clear and strong enough, the sentence might be interpreted in accordance with the contextual and discourse information and be regarded as (at least partially) acceptable.

What we just said about the possibility of resolving conflicts between the product of the first stage and the contextual information brought in at the second stage is perfectly compatible with things working differently when the conflict is internal to the first stage. It does not seem unreasonable to suppose that when there is a conflict in the information that includes just the lexical information provided by the different lexical items and the information provided by the semantic rules, this conflict prevents the creation of a coherent partial interpretation, and the sentence is regarded as unacceptable.

This would explain the complete unacceptability of *I refuted me*. The lexical information associated with *I* determines that the word refers to the speaker, and so does the lexical information associated with *me*. So, it is part of the partial interpretation of the first stage that the two pronouns should be understood as referring to the same individual, whichever the actual world turns out to be. On the other hand, given the structural relationship between the pronouns, disjointness condition (B) prescribes that the referent of the two pronouns can not be assumed
to be the same. The conflict prevents a coherent representation at the first stage, and this makes the sentence unacceptable.

By referring to the two stages as 'first' and 'second' I am, of course, not claiming that this is a temporal description of empirical processes. The first stage is previous to the second just in the logical sense that the second stage presupposes the first, since the partial interpretation which is the product of the first stage is one of the elements used in the second stage to provide the final interpretation. It might very well be that, when an actual evaluation of a sentence takes place, processes corresponding to the second stage take place at the same time as (or even earlier than) processes in the first, in the same way in which semantic processes might take place before the completion of a logically previous process (for instance a syntactic process like, say, determining which was the first word of the sentence uttered --as when we have not clearly heard what was uttered).

Someone might object to my having regarded the evaluation of presuppositions as pertaining to the second stage where contextual information is taking into account. It might be claimed that, since presuppositions are introduced by some specific words or forms of sentences, they should be regarded as part of what is evaluated at the first stage when the lexical information, the structure of the sentence and the semantic rules are taken into account. If presuppositions are evaluated at the first stage then, for many of the examples we have considered, the conflict between different sources of information would already arise at the first stage. So, there would be no justification for my claim that the conflict of information that
explains the (partial) acceptability of sentences like (9b-e) is of a different sort from the conflict of information that makes (34) bad.

We can see that this objection is not correct by taking into account one of the points made by Saul Kripke in Kripke(1990): some presuppositions involve an anaphoric element, analogous to some uses of pronouns. If someone says "Paul also went to the movies" the sentence involves an anaphoric reference to certain proposition, like for instance, that Paul went to the library, or that Arthur went to the movies. We say that some presuppositions involve an anaphoric element in the following sense: the full interpretation of the sentence carrying the presupposition requires the identification of one proposition that has certain characteristics and that is in some way salient. The full interpretation of a referential use of the pronoun he requires the identification of an individual who is male and that is in some way salient (either by having being just mentioned or in some other way). Analogously, the full interpretation of an utterance of the sentence "Paul also went to the movies" requires the identification of a proposition with certain characteristics (it has to involve the attribution of a property to Paul, or one attribution of the property of going to the movies) and which is salient in some way --either by having been recently expressed or in some other way (for instance, by the fact that the speakers are watching a video where they can see Paul going to the library). In the same way in which the determination of the referent of a pronoun is carried out in the second stage of the interpretation where contextual information is brought in, it also belongs to the second stage the determination of what particular proposition is that is being anaphorically referred to by a sentence carrying a presupposition which arises because of the presence of words like also or even, or the stressing
of some word. Similarly for the presupposition that plays a role in interpreting the NPs in the clause \( \beta \) in structures like *What do you mean \( \alpha \)? \( \beta \)*, which also depends on identifying a proposition (the one expressed by \( \alpha \)) that has been previously expressed.

8. Relevant Contexts

8.1 Problematic examples

In this last section we are going to consider new data that will show the necessity of modifying our disjointness conditions so that the context that is relevant for the application of \((B)'\) and \((C)'\) is identified in a more precise way. Consider the following texts

(35) A: That man over there who is talking to Delia is Miguel

B: Right. But Delia does not know that he is Miguel. Actually, she thinks that he hates Miguel.

(36) A and B are talking about their mutual friend David, who is an amnesiac:

A: Has David already realized that he is David?

B: No, this very morning he told me that he saw him in a video

(37) A: I would like to know certain thing about Tim

B: Tim is that man over there. Ask him.

A: I see
B: Even if he were not Tim, you should ask him, because he would be someone playing in the same band (he is wearing the band’s suit) and he would know him well.

These texts seem to pose a problem for the proposal I have been defending (as well, of course, for Standard Binding Theory). In (35) when She thinks that he hates Miguel is uttered A and B are assuming that Miguel is the man they are looking at (B’s saying "Right" helps to indicate that what A has said has been accepted and can be taken as being part of the context). So both he and Miguel pick up Miguel in each world in the context. So the two NPs are presupposed to be co-valued. So according to our disjointness condition (C)' the sentence could not have the interpretation that it actually has. Similar considerations can be made with respect to (B)' and the clauses he saw him in (36), and he would know him well in (37).

8.2. Derived Contexts

I think that we can see what the right way to try to approach the problem is by appealing to the notion of derived context as introduced in Stalnaker(1988).

The point of an speech act is generally to distinguish among different possibilities that at a given moment in a conversation are regarded as open. That is, to distinguish among the possible worlds in the context. Certain parts of the discourse do not primarily distinguish between the possibilities that the participants in the conversation regard as live options but rather between some other set of possibilities.
In a belief attribution, for instance, the clause that expresses what the belief being attributed is does not primarily distinguish among the possibilities that the participants in the conversation regard as open but rather among the possibilities that are open according to what the participants in the conversation take the subject of the belief attribution to believe.\(^3\)

The set of possibilities among which the embedded clause in a belief attribution sentence distinguishes plays a role in regard to the embedded clause similar to the role that the context set plays in regard to the main sentence. Stalnaker calls this set of possibilities a derived context. It will contain all the possible worlds compatible with what it is presupposed in the main context that the subject of the belief attribution believes. We will say that it is the derived context for some particular clause in a sentence, relative to some particular main context.

In the same way that the main context has to satisfy any presupposition carried by a sentence which is uttered in that context, the derived context for an embedded clause has to satisfy the presuppositions carried by that embedded clause: the sentence "John believes that Spei stopped beating her brother" carries the presupposition that John believes that Spei was beating her brother, and so, requires that it be presupposed in the derived context that Spei was beating her brother (it is not necessary that main context has that presupposition).

\(^3\) Of course by distinguishing among the possibilities that are open according to what the participants on the conversation assume that the subject of the belief attribution believes, the embedded clause will indirectly distinguish among the possibilities that are open according to what the participants in the conversation believe (what they assume as open regarding what the subject believes). This is why I say the embedded clause does not primarily distinguish among those possibilities.
In the same way that a referential expression is felicitously used only if there is a suitable individual in each possible world in the context it can pick up, a referential expression in the embedded clause of a belief attribution must pick up a suitable individual in each world in the derived context. I will not try here to give any specific semantics for belief attribution sentences. For our present purposes will suffice to notice that referential expressions in an embedded clause expressing a belief pick up an individual in each world \textit{in the derived context} and so the question can arise if two expressions are presupposed to refer to the same individual in the derived context (i.e. whether they pick up the same individual in each world in the derived context).

We have seen exemplified in the case of the derived contexts for the embedded clause of belief attributions three characteristics that identify derived contexts in general: first, a derived context is the set of possibilities the corresponding clause distinguish among; second, a derived context must satisfy the presuppositions of the corresponding clause; third, in each world in the derived context there must be a referent for each referential expression in the corresponding clause.

Subjunctive conditionals (as well as indicative conditionals) are also evaluated in part with respect to a derived context: the consequent of the conditional is evaluated with respect to a set of possibilities similar in some relevant sense to the possibilities which are open in the main context, but such that what the antecedent says is true. The presuppositions of the consequent clause must be satisfied in the derived context (as it happens in "If Spei had been beating her brother, she would have stopped beating him by the time he was 33"). A referential expression
in the consequent of a subjunctive conditional will pick up an individual in each world in its derived context (as it happens in "If there were a thief on the roof, he would be very cold").

8.3. Disjointness Conditions

I think we will be able to account for sentences like (35)-(37) by modifying the Disjointness Conditions so that they take into consideration the possible existence of different contexts in the evaluation of certain parts of a sentence. They could be modified in the following way:

(B)* If a sentence whose LF representation is of the form ...α...β... (where β is a pronominal, and α is an NP that c-commands β in its GC) is used in a context C, and C' is the relevant context relative to C for a clause containing α and β, then it is not presupposed in C' that α and β are co-valued.

(C)* If a sentence whose LF representation is of the form ...α...β... (where β is an R-expression, and α is an NP that c-commands β) is used in a context C, and C' is the relevant context relative to C for a clause containing α and β, then it is not presupposed in C' that α and β are co-valued.

Regarding the notion of the 'relevant context' we stipulate that if C' is the derived context for a clause s' of a sentence s uttered in C, then C' is the relevant context relative to C for s'. For

-76-
If we consider (35) and (37) with respect to (B)* and (C)* (we will comment on (36) in the next section), we can see that they do no longer present a counter-example to the our disjointness conditions.

The clause *He hates Miguel* in (35) is part of a belief attribution sentence. The two NPs would pick up the same individual in each world in the main context, but they pick up different individuals in the derived context for this clause. According to what A and B suppose that Delia believes, Miguel and the man that Delia is talking to are two different people (or at least it is not determined that they are the same --i.e. at least for some world in the derived context, they are not the same). So the two NPs in *He hates Miguel* are not presupposed to pick up the same individual in each world in the relevant context, namely, the derived context for the embedded clause of the belief attribution sentence.

An analogous explanation applies to the clause *he would know him well* in (37): in the relevant context, i.e., in the derived context for the consequent of the subjunctive conditional, the man A and B are looking at is not Tim, since in all the worlds in the derived context for a subjunctive conditional what the antecedent says is true, and the antecedent of the subjunctive conditional in (37) says that the man A and B are looking at is not Tim. So, if for each world in the derived context *he* picks up the man A and B are looking at, and *him* picks up Tim, the
two NPs are not presupposed to be co-valued (actually they are presupposed not to be co-valued, i.e., they pick up a different individual in each world in the derived context).

8.4. 'Saying Attributions'

In the case of sentences containing what we might call a saying attribution it does not seem possible to specify a 'derived context' that plays all the roles that we have seen a derived context plays for belief attribution sentences and subjunctive conditionals.

One important difference between belief attributions and saying attributions is that the former are related to a state whereas the latter are related to an event: saying attributions describe the content of some particular utterance that the subject of the saying-tribution performed. Even when we use several saying attribution sentences about the same subject, there might not be one single set of possibilities which each saying attribution makes a contribution in trying to specify. Each saying attribution might be about a different utterance, and it will be concerned with specifying which worlds are compatible with the content of that particular utterance. This is not the same as specifying the possibilities left open by all what the subject said. I might report that John said that he is from Cambridge, and that he said that he is from Newport. The two saying attributions do not attempt to describe a single set of possibilities --the one compatible with what John 'said', but rather they report two different utterances of John, each with its own content.

We could still say that the derived context for a saying attribution is the set of possible worlds compatible with what we believe is the content of the utterance we are attributing to the
subject when we say that he said so and so. If this were the derived context then it would be
unlike the derived context for belief attributions or conditionals, since it does not have to contain
the presuppositions that the words used in the saying attribution might carry: as it is well known,
sentences with 'verbs of saying' (say, tell, announce, ask) are such that the sentence containing
them does not require any presupposition that might usually be carried by some of the
expressions in the embedded clause. So, for instance, I can felicitously and even truly say of
my bachelor friend Ambròs that he said that his wife had stopped beating the king of France.

It seems that referential expressions in the embedded clause of a saying attribution
sentence must have a suitable referent in the following set of possible worlds: those worlds that
according to the main context are compatible with what the subject of the saying attribution was
assuming when making his utterance. So, for instance, I can utter (38)

(38) Gemma believed that some moonlight reflected on the room wall was a ghost. She said
that he probably was friendly.

When making the saying attribution we are assuming that in all the worlds compatible with what
Gemma was assuming when she made her utterance there was a ghost in the room. He in the
embedded clause will pick up, for each world in the set of worlds compatible with what Gemma

---

34 See, for instance, Karttunen (1974).
was assuming when she made her utterance, the ghost that Gemma thought was in the room. I propose that we regard the set of worlds that in the main context are regarded as compatible with what the subject of the saying attribution was assuming when he made the utterance we are attributing to him as the relevant context for the embedded clause of a saying attribution sentence.

Then we can account for (36): the relevant context for the embedded clause in *he told me that he saw him in a video* is the set of worlds compatible with what A and B assume that David was assuming when he made the utterance we are attributing to him. From what is said in the first part of (36) it is clear that in each possible world compatible with what A and B know that David believed and that he believed that people who were listening to him believed and, so, in each world compatible with what they know that David assumed when he uttered a sentence there are two different people: himself and David. So the pronouns *he* and *him* do not pick up the same individual in the relevant context for the embedded clause, and so the interpretation of this clause in not in conflict with our new version of the disjointness conditions.

8.5. Checking our examples

As pointed out in footnote 16, the use of certain expressions (such as "this is why...") can alter what is the context which is relevant for evaluating some sentence. These expressions might

---

35 Notice that *he* would not have a referent in each world in the main context, and this is why we could not say: "Gemma believed that some moonlight on the room’s wall was a ghost. He probably was friendly".
allow a sentence to have certain interpretation that the disjointness conditions would not allow the corresponding sentence without these expressions to have. Also it is easy to construct a sentence that has an interpretation that seems not to accord with what the disjointness conditions prescribe by using definite descriptions or proper names that have a very salient associated description (such as the usual "Hesperus", "Phosphorus", "Superman", "Clark Kent"): when interpreting a sentence the hearers always try to accommodate what they are assuming so as to make the sentence interpretable; if, for instance, two co-referential proper names each has an associated description, then, if necessary, it is easy to imagine how a world would be like where the two names were not co-referential, and to assume that the sentence should be evaluated with respect to that kind of worlds. In summary: there are several devices that might change the context where a sentence is evaluated; in order to make sure that what explains (35)-(37) is what we say that explains them (i.e. that certain parts of the sentences must be evaluated with respect to a context which is not the main context) and not something else, we should do the following: we should check that the relevant clauses can not be interpreted as in (35)-(37) when we make just the minimal changes necessary to prevent evaluating the clause with respect to a derived or special context (i.e. different from the main context).

Consider (35)*, (36)* and (37)*

(35)* A: That man over there who is talking to Delia is Miguel

B: Right. But Delia does not know that he is Miguel. Actually he hates Miguel.
A and B are talking about their mutual friend David who is amnesiac:

A: Has David already realized that he is David?
B: No, this very morning he saw him in a video.

A: I would like to know certain thing about Tim
B: Tim is that man over there. Ask him.
A: I see.
B: You should ask him, because he is someone playing in the same band (he is wearing the band’s suit) and he knows him well.

The clause *He admires Miguel* in (35)* is no longer acceptable if we intend *He* to pick up, for each world in the context, the man that A and B are looking at and for *Miguel* to pick up the man A and B call "Miguel". Similarly for *he saw him in a video* in (36)*, and *he knows him well* in (37)*.
References


Chapter Two

On the Interpretation of Second-order Quantification:

An Examination of Some Criticisms of Boolos’ Proposal

George Boolos offers an innovative proposal of how to understand second order quantification in two papers: To be is to be the value of a variable (or to be some values of some variables) and Nominalistic Platonism¹. Many times it is difficult to provide much positive evidence for a proposal that, like Boolos’, concerns very basic ideas and intuitions. We will see that Boolos nevertheless provides some positive evidence. One good way of trying to make one such basic proposal plausible, though, is by showing that there is no negative evidence against it. This paper is mainly concerned with this second sort of enterprise. Several authors have offered criticisms to Boolos’ proposal. In this paper I will examine the criticisms that have been put forward by Charles Parsons, Patrick Grim and Michael Resnik. The criticisms offered by these authors cover different aspects in Boolos’ argumentation. In the first section of this paper I will briefly summarize Boolos’ proposal. In the following three sections I will examine the criticisms offered by each of these authors.

¹ To be is to be the value of a variable (or some values of some variables), The Journal of Philosophy 81, August 1984, pp. 430-449. Nominalistic Platonism, Philosophical Review 94, July 1985, pp. 327-344. Both papers are collected in George Boolos’ Logic, logic and logic, Harvard University Press, forthcoming.
I will refer to *To be is to be the value of a variable (or to be some values of some variables)* and *Nominalistic Platonism* as TB and NP, respectively.

1. **Boolos’ proposal**

1.1 **The main ideas**

There are some propositions, and, in particular, there are some propositions about set theory, that we regard as true, which are not expressible by means of a first order formula but that would seem to be expressible by means of a second order formula. Some examples of them are the separation principle and set theoretic induction. These principles would seem to be formalizable respectively as

(a) $\forall X \forall x \exists y \forall z (z \in y \leftrightarrow z \in x \land Xz)$

(b) $\forall X( \exists x Xx \rightarrow \exists x[ Xx \land \forall y (y \in x \rightarrow \neg Xy)] )$

Standardly, second-order variables are interpreted as ranging over classes of the objects which are the values of the first order variables. Then, though, consider for instance, the separation principle: given some sets (for example, the sets that have only one element) and given a set $x$ there is a set that has as elements all the given sets that are elements of $x$. The full expression of this principle under the standard way of interpreting second order quantification
requires that there are classes that have as elements, for instance, all the singleton sets. If there were such classes they would not themselves be sets (they are usually called 'proper classes'). So, under the standard interpretation of second-order quantification, if we want a sentence such as (a) to fully express the separation principle we believe we have to assume that there are classes which are not sets. This is completely undesirable if we think, as Boolos does, that set theory is supposed to be a theory about all the class-like objects that there are.

Another way of interpreting second other sentences which would not involve accepting proper classes, would be to require that the objects over which the individual variables in the second order formula range form a set; but then we would not be able to fully express principles like the ones above.

George Boolos argues that sure enough we can fully express the principles we believe (like separation and the induction principle) and that we can express them by means of second order formulae. Interpreting second order formulae does not require that it be assumed that second order variables range over classes of the individuals which are the values of the first order variables. This is so because:

(1) Second order formulae can be interpreted by appealing to natural language plurals, and

(2) This use of natural language plurals does not imply commitment to the existence of classes of the individuals in the domain of discourse.
In the next two sub-sections I will examine the arguments given for (1) and (2) in TB and NP.

1.2 Interpreting second-order formulae by appealing to plurals.

Nonfirstorderizable sentences. In both articles (TB 431-440, NP 327-329) there is a discussion of different natural language sentences and arguments and the possibility or impossibility of symbolizing them in a first-order language. These examples serve to show the expressive power of natural language plurals (sometimes a sentence with plural quantification can not be expressed in first-order logic, often its most 'natural' symbolization is by means of a second-order sentence). The discussion of these examples evince a correspondence between second-order quantification and plurals in one direction: many sentences containing plurals are most naturally symbolized in second-order logic. What we want to justify, though, is the other direction: second-order sentences are interpretable by means of natural language sentences containing plurals. The way in which the discussion of non-firstorderizable sentences helps to justify this seems to be the following: by showing that several natural language sentences containing plurals are most naturally formalized by means of second order formulae we also show that the second-order formulae that formalize the English sentences containing plurals that we have considered are themselves, in turn, naturally interpreted in English by a sentence containing plurals; this fact would suggest that, unless we have any reason to suppose that the second-order sentences that we have seen formalize some English sentences containing plurals have some structural feature
or other characteristics that distinguish them from other second-order sentences (and we do not have any reason to think that this is so), these other second order sentences will also be naturally interpretable by means of sentences containing plurals. The discussion of non-firstorderizable sentences helps to give some intuitive plausibility to the idea that we can satisfactorily interpret second order formulas by means of natural language plurals. This discussion does not constitute by itself, though, a justification of (1).

The claim in (1) is shown to be true in two ways\(^2\): (i) first by providing a translation procedure between a second order language and (an augmented version of) English; (ii) second, by actually providing a truth theory for a second order language. This second task is accomplished in a two-fold manner: on the one hand, a truth theory using (an augmented) second order language is provided; on the other hand, the translation procedure mentioned in (i) makes it possible to understand the second order language that it used as a metalanguage for giving the truth theory.

**Translation.** We can consider an augmented version of English in which pronouns (including the relative pronoun "that") have subindexes and in which coindexation means co-valueness (i.e., in the case of referential expressions it means co-reference, in the case of expressions bound by a quantificational noun phrase, the index indicates which expressions are bound by a given quantificational noun phrase). This augmented English seems to be as intelligible as English

\(^2\) For a general discussion of how these two ways of *interpreting* a language differ, see Chapter 3.
itself. Boolos offers (in TB p. 444) a procedure to translate any second order formula of the language of set theory into this augmented English. Natural language plural quantification is used to translate second-order quantification. The procedure is the following:

- \( \forall v \) translates as *it, is one of them* \\
- \( v \in u \) translates as *it, is a member of it* \\
- \( v = u \) translates as *it, is identical with it* \\
- \( \& \) translates as *and* \\
- \( \neg \) translates as *not* \\
- \( \exists v F \) translates as *there is a set that, is such that* \( F' \) (where \( F' \) is the translation of \( F \)) \\
- \( \exists VF \) translates as *either there are some sets that, are such that* \( F' \), or \( F'' \) 
  (where \( F'' \) is the result of substituting an occurrence of \( \neg v = v \) for each occurrence of \( \forall v \) in \( F \))

Notice that if we allow not only pronouns but all noun phrases to have coindexation, then we can add clauses that translate universal quantification in the following way:

- \( \forall v F \) translates as *given a set, \( F' \)* (or also, *whatever a set, is, \( F' \)*) \\
- \( \forall VF \) translates as *given some sets, \( F' \) and \( F'' \)* (or also, *whatever some sets, are, \( F' \) and \( F'' \)*)

Notice that this translation procedure does not give us an explicit account of how the semantic value of complex expressions depends on the semantic value of simpler parts. This is provided by the truth theory:
The truth theory. The standard view is that the use of first-order sentences carries a commitment to the objects in the domain the sentence is taken to be about, and that the use of second-order sentences carries a commitment to all the classes of the objects in the domain. It would seem that the basis for claiming this is that in giving a truth theory for a language and, in particular, in giving a definition of satisfaction of a formula by a sequence, first-order variables are assigned objects in the domain and second-order variables are assigned classes of those objects. If this is the only basis for claiming the ontological commitment of second-order sentences to classes, then Boolos refutes it by providing a truth theory for a second-order language in which there is no assignment of any entity to second-order variables. The language he chooses is among the ones that would seem more difficult to deal with: the second order language of set theory (since on the intended interpretation, the domain is not a set). He defines the satisfaction predicate: \( R \text{ and } s \text{ satisfy } F \), where \( s \) and \( F \) are first order variables, and \( R \) is a second-order variable. A formula \( F \) will be true if there are a sequence \( s \) and some sets \( R \) that satisfy it. The recursive definition of satisfaction is the following (where "<,>" is the ordered pair function sign):

If \( F \) is \( u \in v \), then \( R \text{ and } s \text{ satisfy } F \) iff \( s(u) \in s(v) \)

If \( F \) is \( u = v \), then \( R \text{ and } s \text{ satisfy } F \) iff \( s(u) = s(v) \)

If \( F \) is \( \forall v \), then \( R \text{ and } s \text{ satisfy } F \) iff \( R < V, s(v) > \)

If \( F \) is \( \neg G \), then \( R \text{ and } s \text{ satisfy } F \) iff \( \neg (R \text{ and } s \text{ satisfy } G) \)

If \( F \) is \( (G \& H) \), then \( R \text{ and } s \text{ satisfy } F \) iff \( (R \text{ and } s \text{ satisfy } G \& R \text{ and } s \text{ satisfy } H) \)

If \( F \) is \( \exists v G \), then \( R \text{ and } s \text{ satisfy } F \) iff \( \exists x \exists t (t \text{ is a sequence } \& t(v) = x \& \forall u (u \text{ is a first-order variable } \& u \neq v \rightarrow t(u) = s(u)) \& R \text{ and } t \text{ satisfy } G \) )
If $F$ is $\exists VG$, then $R$ and $s$ satisfy $F$ iff
\[
\exists X \exists T \left( \forall x( Xx \leftrightarrow T<\forall V,x>) \right) \& \forall U(U \text{ is a second-order variable } \& U \neq V \rightarrow \forall x(T<U,x> \leftrightarrow R<U,x>)) \& T \text{ and } s \text{ satisfy } G
\]

The last two clauses could be simplified in the following way:

If $F$ is $\exists vG$, then $R$ and $s$ satisfy $F$ iff
\[
\exists t \left( t \text{ is a sequence } \& \forall u(u \text{ is a first-order variable } \& u \neq v \rightarrow t(u) = s(u) ) \& R \text{ and } t \text{ satisfy } G \right)
\]

If $F$ is $\exists VG$, then $R$ and $s$ satisfy $F$ iff
\[
\exists T(\forall U(U \text{ is a second-order variable } \& U \neq V \rightarrow \forall x(T<U,x> \leftrightarrow R<U,x>)) \& T \text{ and } s \text{ satisfy } G )
\]

Three comments: (a) Notice that the truth theory is provided in the very second order language of set theory (the object language) with the only addition of the predicate '$R \text{ and } s \text{ satisfy } F$'; in particular, notice that all the other predicates appearing in the definition (like 'is a sequence' or 'is a first order variable') can be expressed in the language of set theory: for instance, 'is a variable' can be seen as a condition on sets since we can regard variables as (Gödel) numbers, and numbers as sets. All the appearances of variables written using capital letters for which I have used a boldface font are not genuine second-order variables of the language in which we are providing the truth theory, but rather they stand either for a first order variable or for the functional expression that gives (the set corresponding to) the Gödel number of the second-order variable involved. (b) This truth theory provides us with what we could ask from it: from this theory plus standard axiomatic second-order logic we can derive the usual
lemmas about free variables, and laws of truth; if, after all, the one who claims that we should interpret second-order variables as having classes of the objects in the domain as values were right, this theory does not alter the truth values assigned to the sentences. (c) Since the theory is given in the (augmented) second-order language of set theory, which is not a language we know, we need the translation procedure to interpret the clauses in the definition of satisfaction. Combining the translation procedure and the truth theory we have a set of sentences in a language we understand (augmented English) which tells us in the appropriate way how the satisfaction conditions of complex expressions depend on those of their parts.

So the general picture is this: we can interpret second order quantification without having to appeal the classes of the objects in the domain; this is so because we can interpret second order quantification appealing to natural language plurals. Now, the proposal would not work if the use plural quantification in natural language were just a covert way of quantifying over classes of the individuals in the domain of discourse.

1.3 On the ontological commitments of natural language plurals

I think that it is hard, in general, to give an argument to show that a sentence does not have certain ontological commitment beyond appealing to the intuition that it does not. Suppose that someone claimed that in saying 'Bill runs' we are referring to a set that has all the individuals in the domain of discourse but Bill as its members; how would we argue against this claim, besides appealing to our clear intuition that there is no such commitment? It seems that the
burden of proof should be on the one that claims that there is some specific ontological commitment.

I think that, in addition to appealing to the intuitive apparent lack of commitment to classes of all the objects in the domain in the use of some sentences containing plurals, there are basically three kinds of considerations in TB and NP trying to defend the view that plural quantification does not carry a commitment to classes. The first two are aimed at showing that some arguments that the defender of the 'commitment to classes' view could put forward are not good, only the last one is aimed at giving positive evidence that there is not such commitment. They are the following:

(a) Someone might claim that sentences with plural quantification are intelligible only because we interpret quantification such as 'there are some dogs...' as 'there is a class of dogs...'. Someone might say that because she thinks that we are only going to be able to give a semantic theory of English if we assign semantic values to expressions using set theory: singular referential expressions may have as a semantic value individuals in the domain of discourse, but plural expressions will have to have as a semantic value classes (or may be classes of classes, etc) of the objects in the domain.

Boolos replies by claiming that there is a distinction between (i) what (parts of) languages we can understand (ii) what (parts of) languages we are able to give a systematic theory for. We might only be able to give a systematic theory of a kind of 'plural quantification' that works as
quantification over classes, this does not mean that plural quantification in English has to work that way or that we could only understand it if it worked that way.

Whatever the plausibility of Boolos’ reply, it seems that he is granting something that he does not need to: that any adequate systematic semantic theory will be stated in terms of set theory. This is not true in general, and we do not have any reason to think it true for the specific case of a theory of plural reference and plural quantification in natural language. Even if it is true in general that: if any adequate theory for certain expression must assign certain sort of semantic value to that expression then that means that the use of that expression involves that sort of semantic value, we do not need to believe that the use of plurals involves an appeal to classes of the individuals in the domain, since we have not been given any compelling reason for believing that an adequate systematic theory of plurals can only be given if their semantic value is given in terms of classes.

(b) Someone could claim, following Quine, that to determine what the ontological commitment of certain sentence is one has to formalize it using first order logic. Then the sentence conveys commitment to any entity that the variables could take as a value. So, for instance, in the case of ’there are some dogs that are such that one of them leads the others and no one else’ we would have

\[
\exists c ( c \text{ is a class} \ \& \ \forall z (z \in c \rightarrow z \text{ is a dog}) \ \& \ \exists y \exists z (y \in c \ \& \ z \in c \ \& \ y \neq z \ \& \ \forall x (x \in c \ \& \ x \neq y \rightarrow y \ leads \ x) \ \& \ \forall x (y \ leads \ x \rightarrow x \in c) )
\]
and so, it would be determined that the sentence conveys commitment to a class of dogs.

There are, though, some examples of sentences which clearly do not seem to have any commitment to classes but that, if we applied the translating-into-first-order criterion would turn out to have such commitment. Two such examples would be 'For infinitely many stars, there is no planet around them’ and 'Most people like bananas’. So the criterion is not right in general. So one can not use the general validity of the criterion to argue that some plural sentences carry a commitment to certain classes.

(c) A sentence like 'There are some sets that are self-identical, and every set that it is not a member of itself is one of them’ says something trivially true, but if it conveyed a commitment to classes then it would be equivalent to 'there is a collection of sets that are self identical and every set that it is not a member of itself is a member of this collection’ and this sentence is false. So the first sentence can not convey a commitment to classes. (To make it easy to refer to this argument later on, I will call it ARG).

It seems that the defender of the commitment to classes view could reply that the sentence does entail the existence of classes and contra what we could at first think the sentence is false.

The answer given to this reply (in TB-447) is that it is obvious that in general a sentence of the form "There are some As of which every B is one" is synonymous with the corresponding
sentence of the form 'There are some As and every B is an A'. If there is no commitment to a class in the latter, there is no commitment to a class in the former. And it seems harder to claim with respect to the latter that it involves a commitment to the class of all the As or of all the Bs.

In a similar argument (in NP 331-332) Boolos claims that the following three sentences (a)-(c) say the same:

(a) There are some classes such that every infinite class is one of them.
(b) There are some classes and every infinite class is one of them.
(c) There are some classes and every infinite class is a class.

If there is no commitment to a class containing all infinite classes in (c), then there is no such a commitment in (a) either.
2. Charles Parsons’ criticisms

In section seven in *The Structuralist View of Mathematical Objects* Charles Parsons is concerned with whether, when pursuing an eliminative structuralist program, recourse to second-order logic can be fruitful in view of being able to eliminate mathematical objects. He considers two different ways of interpreting it; one of them is Boolos’. Parsons concludes that "Boolos has ... not made a convincing case for the claim that his interpretation of second-order logic is ontologically noncommittal." (p. 328). We will examine what I think are the three main criticisms Parsons presents against Boolos’ proposal. Before doing that, though, I would like to point out the following:

Parsons concentrates his analysis/criticism of Boolos’ proposal on the truth theory for the second-order set-theory language that Boolos offers in NP. Parsons considers the question of the ontological commitment of natural language plurals only in a couple of paragraphs (pp. 326-327). He seems to conclude (though he is not very explicit about it) that there is no clear answer to the issue of the ontological commitment of natural language plurals and that this creates a ‘discomfort’. He seems to suggest that Boolos offers his theory of truth to deal with this discomfort with natural language plurals. But this is not so. According to the general structure of Boolos’ argumentation that we have presented in section 1, the truth theory that Boolos offers does not directly show anything about the ontological commitment of natural language plurals. It is the truth theory that gets its point from the fact that natural language plurals carry no

---

3 Syntheses 84, 1990, pp. 303-346.
commitment to classes, not the other way around.\(^4\)

2.1 First objection

Boolos claims that in the truth theory for the second-order language of set theory that he gives in \(NP\) he makes no use of an assignment of a value to second-order variables. Parsons says that "it is hard to agree with Boolos that the treatment of second-order variables in the definition does not offer scope for the notion of a value comparable to what it offers to the notion of a value of an individual variable" (p. 328). We could say that the value of the second-order variable \(V\) is \(\lambda x R(<V, x>)\), that is, the class that has as a member any set \(a\) such that the pairing of (the set that corresponds to the number that is the Gödel number of the variable) \(V\) and \(a\) is one of the \(R\)'s. Parsons claims that the \(R\)'s are simply coding an assignment of classes to second order-variables, and he suggests that Boolos' truth theory does not essentially differ from the standard one in regard to the use of classes to interpret second-order quantification.

We can agree with Parsons when he says that we can see the \(R\)'s as coding an assignment of classes to values and also vice versa, only if we assume that given some sets there is going to be a (maybe proper) class that has exactly those sets as elements. If we take set theory to be a theory about all the class-like objects, though, we do not need to concede that assumption. (Even if intuitions would suggest that given some objects there must always be one thing which

\(^4\) The truth theory might be taken to have the following indirect relevance to the issue of the ontological commitments of natural language plurals: if second order logic could only be interpreted by appealing to classes then it would seem that any formula that can not be formalized in a first-order language and can only be formalized in a second-order language should itself involve reference to classes.

-101-
is constituted by them, the set theoretic paradoxes teach us not to follow these intuitions). If we accept it, though, then it will certainly be the case that given some R’s (i.e. some pairs where the first member of the pair is a second order variable) a unique assignment of classes to second-order variables is determined, and given one such assignment some specific sets are determined.

Given some sets, the R’s, which are such that each set is an ordered pair such that the first member is (the set corresponding to the Gödel number of) a variable, we can define the corresponding assignment function, which we may call $f$, in the following way: for each second-order variable $V$,

$$f(V) = \text{the class that has as member any set } a \text{ such that } <V,a> \text{ is one of the R’s}$$

And given an assignment $f$ of classes to second-order variables we can consider the pairs such that the first member of the pair is a second order variable and the second member is an element of the class that $f$ assigns to the first member of the pair.

Having said that, I do not think that Parsons is right in suggesting that from the fact that we can see the R’s as coding an assignment, it follows that when we give the truth theory and we refer to the R’s, we are referring to the corresponding assignment and the values that the variables would have according to that assignment. From the fact that the R’s can be seen as an coding an assignment, it does not follow that they are one. There is a difference between what information we can obtain from something, and what that thing is. For instance, we can see every
object as coding a certain ordered pair: the one that has that object as its first member and
Archibald Leach as its second; and vice versa, we can see each ordered pair of the form
<x,Archibald Leach> as coding an object: the one that is its first member. This does not mean
that talk of one is like talk of the other.

2.2 Second objection
Parsons proposes "an ontological intuition" which is "a little different from but complementing
Quine's" (p. 328), and consists in thinking that the ontological commitment of a natural language
sentence is carried by the expression that plays the role of subject. The subject indicates what
we are talking about, and so what we are committing ourselves to. In a formal language, the
corresponding role to be the subject of a sentence is to be an argument of a predicate. So, a
sentence in a formal language conveys commitment to those entities that can be the interpretation
of the arguments of a predicate. In the truth theory in NP the predicate 'satisfy' has three
arguments whose interpretation will be a sequence, some pairs, and a formula. So, some pairs,
let's call them the R's, can be one of the arguments, and so they constitute a unity (whether we
see it as a class or as a plurality). So, in including a predicate, satisfy, that takes a second-order
argument, the truth theory in NP involves commitment to second-order entities like classes,
collections or pluralities.

Leaving aside some other probably minor objections to Parsons ontological criterion (Not
only the subject of a natural language sentence might be an argument of a predicate in its
formalization, but also the object and some other complements; Are we committing ourselves to ghosts if we say "Ghosts do not exist"?) I think that the following can be replied: There are clear cases in which a predicate taking a second-order argument does not convey commitment to classes. Consider "Joe and Jason met after class". It seems clear that this sentence does not include any reference to a set, a collection or a plurality (if that is a single entity made out of many), but only to Joe and to Jason. The same is true of "The students met after class", which is a sentence where the predicate takes and argument, the students, which is of the same kind as the argument the R’s in the R’s and s satisfy F. [If someone replies that even if there is no reference to a class in "Joe and Jason met" there might be one in "The students met", we can ask her to consider the following sentences and to tell us in which of them the reference to a class comes into: "Joe and Jason met", "Joe, Jason and Daniel met", "Joe, Jason and the other student met", "Joe and the other two students met", "Joe and the other ten students met", "The ten students met", "The students met"].

It might be replied to what I just said that the predicate 'satisfy' in the truth theory is different from 'meet' in that it does not have a single (plural) argument as 'met', but three arguments (of different kind) and that it is this fact that might make the R’s be ‘a unity’ which is different from the other two arguments. To this I would answer that we can also consider a sentence like "Johnny put the pieces together" (or if you want, "Johnny put piece A and piece B together", "Johnny put two pieces together", "Johnny put four pieces together", "Johnny put more than three pieces together", "Johnny put some pieces together", "Johnny put the pieces together"). If we formalize this example the predicate will take two arguments, a first-order and
a second-order one. Nevertheless it seems to me that we have intuitions as clear as in the examples in the previous paragraph that this example does not include reference to a class, but only reference to, on the one hand, one individual, John, and on the other hand, to some other individuals, the pieces he put together.

2.3 Third objection

In a footnote (footnote 64, p.344) Parsons observes that, as we mentioned above, in the truth theory offered in NP, R is an argument of the predicate 'satisfy' and he claims that there is a difficulty for Boolos’ proposal to express that predication using plurals. He expresses in English the last clause in the recursive definition of satisfaction \((R \text{ and } s \text{ satisfy } \exists V G)\) in accordance to Boolos proposal of interpreting second-order variables in terms of plurals. Simplifying the formulation Parsons gives we could express it as

There are some variable-pairs \(the T's\) such that for any second order variable \(U\) distinct from \(V\), the pairing of a set and \(U\) is one of \(the T's\) if and only if it is one of \(the R's\), and \(the T's\) and \(s\) satisfy \(G\)

Parsons claims that there is a difficulty in the expression "the T’s and s satisfy G", since nothing marks the final "the T’s" as a second order argument; "the T’s and s satisfy G" could say that all or most of the T’s are x’s such that x and s satisfy G.

(Notice that this objection is a kind of contrapositive of the previous objection: We can
take the present objection to be 'If the R's are not a single unity then there is no way of identifying them as an argument', whereas the former was 'since the R's are an argument they have to be a unity').

I would reply the following. First, in regard to the remark that we could interpret "The T's and s satisfy G" so as to be saying something only about most of the T's: It is true that we sometimes say "the A’s P" when we know it is true only that "most of the A’s P"; but this is just a pragmatic phenomenon: on the one hand, it seems that "the A’s P" expresses a simpler thought than "Most of the A’s P"; on the other hand, when not all the A’s P, "most of the A’s P" is more informative than "The A’s P"; which expression we use in a certain circumstance will depend on a trade-off between on the one hand, quantity of information and on the other relevance and simplicity. One characteristic of the mathematical use of language is that it is fully explicit and does not rely on pragmatic mechanisms. When using the expression "the T’s" in giving the truth theory we have to assume that the pragmatic balance mentioned above does not take place and that the expression has to be understood in its most literal sense.

Second, I do not think that "the T’s and s satisfy G" can be read as 'for each x that is a T, x and s satisfy G’ , in the same way that "Kathrin and her sisters carried the pianos upstairs" does not have (even marginally) a reading where the sentence is true only if for each of Kathrin’s sisters there is an event of carrying a piano upstairs such that she and Kathrin (and no one else) are the agents.
I think, nevertheless, that there is, in principle, a possible reading of the expression "the T’s and s satisfy G" that Parsons does not consider but that does seem to posit a problem for claiming that we can interpret the truth theory in terms of plurals. It is the following: it seems that "the T’s and s satisfy G" could mean that each of the T’s satisfy G and that s satisfy G. It could be claimed, then, that this undesirable reading can only be ruled out if the T’s constitute an unity, and so that we need to appeal to class-like objects in giving the truth theory. To this I would reply the following.

There is, in general, a distinction between the so called collective predicates and distributive predicates. 'Gather', 'disperse', 'love each other' and 'to be 12' are examples of collective predicates; 'sleep' and 'to be red' are examples of distributive predicates. Properly speaking the distinction is not between predicates but between readings of a predicate, since many predicates can have both readings. The classic example is "Ten boys carried a piano upstairs" which in the distributive reading implies that at least ten events of carrying a piano upstairs took place, but it does not in the collective reading. The truth of a sentence with a plural subject and a distributive (reading of a) predicate will imply the truth of each sentence that has the same predicate, and has as subject an expression referring to one of the individuals that the plural subject refers to; this is in general not true of sentence with collective predicates.

The objection under consideration requires that in "The T’s and s satisfy G", 'satisfy' has a distributive reading over the T’s and s, if we take "the T’s" to refer to many individuals and not to a single class or plurality.
The predicate 'satisfy' in "The T's and s satisfy G" does not have such a reading. What the interpretation of "satisfy" is depends on how we define it. Consider the other clauses in the recursive definition. If 'satisfy' allowed a distributive reading over the T's and s, then it should be defined, for instance, what it means that one of the T's satisfies \( \mu \in \nu \), or that s satisfies \( \forall \nu \); but this is not defined.

2.4 Is there commitment to classes in some examples involving plurals?

There are some plural expressions that at least in some uses do seem to refer to 'collective unities'. Consider, for instance, "The Klinons are aggressive". Someone might believe what this sentence expresses because, say, of what she has learned about the actions of the Klinon army, and she might believe it even if she does not believe that all or most of the Klinons are aggressive. This would suggest that her belief is not about some persons, the Klinons, but rather about some one entity whose properties depend in some complex way on the properties of each individual who is Klinon. Also we would say that the sentence "The Rolling Stones are in the room" is false if, for instance, Ron Wood is not in the room, but at the same time would maintain that "The Rolling Stones got their first hit in 1965" is true. So, it seems that "The Rolling Stones" is not a term that simply refers to some individuals.

I am not clear about how exactly these examples should be understood, or what

5 The Rolling Stones original lead guitarist, Brian Jones left the group soon before he died in 1969. He was then replaced by Mick Taylor, who in turn was later replaced by Ron Wood.
implications they have. I wish I were.

In any case they do not seem to affect Boolos' proposal. Boolos does not need to claim that NO use of natural language plurals conveys commitment to a class or collection of the objects in the domain, but rather that certain uses of plurals (the ones he proposes to use to interpret second-order formulas) do not convey such a commitment. Also, Boolos does not claim that ALL uses of plurals are clear enough as to use them in our logic theory, but only that certain uses are clear enough and that they suffice to interpret second-order languages. Similar remarks would apply to the case of first-order logic. When giving an interpretation for a first-order language we use natural language singular quantification and singular terms. This does not mean, though, that we understand well enough any use of singular quantification or of referring expressions in natural language as to base our logic on them. For instance, this does not mean that we are clear enough about all the uses of definite descriptions. It suffices that the very restrictive part of natural language that we use in giving the interpretation of first-order languages is clear enough.

Do we understand quantification such as "there are some objects that are such that..." as clearly as we understand "there is an object such that ..."? I do not know what exactly would be required to answer the question one way or the other. And, in any case, I will not pursue this issue any further here.
3. Patrick Grim’s criticisms

At a certain point in *The Incomplete Universe: totality, knowledge, and truth* Patrick Grim is concerned about using a formal language in which he intends to express quantification over all propositions. He claims to have shown, though, that there is no set having as members all the propositions. Nevertheless, 'the only formal semantics for quantification we have is in terms of sets' (p.115). It is at this point that he considers abandoning formal semantics, and when he comments in a very long note on what, according to him, is one attempt of doing so: Boolos’ proposal in TB and NP. According to him Boolos’ proposal of interpreting second order quantifiers by appealing to natural language plurals does not offer any "genuine escape" from the "problems that motivated it". Let’s explain and examine Grim’s argumentations.

Here are Grim’s main assertions:

(1) According to Grim, what Boolos proposes is to substitute the formal semantics for second order languages in terms of sets or classes or collections for an informal semantics in terms of natural language plurals.

(2) The aim of this is to "escape from the Russellian problems" which "arise in set-theoretic


\[^{7}\] All the quotations will be from page 152, unless otherwise indicated.
semantics when we quantify over sets" and "in class-theoretic semantics when we quantify over classes".

(3) Boolos' proposal does not succeed, though, since it falls into a similar problem. This can be shown in the following way:

(3.1) Let's consider a "language rich enough to quantify over plural noun phrases and to express some of their basic properties. Consider, for instance the following second-order formula:"

\[
\exists X \ ( \exists x X x & \forall x [X x \leftrightarrow \neg A x x] )
\]

(3.2) "'x' is taken as ranging over plural noun phrases and '¬Axx' is taken to indicate that a noun phrase does not apply to itself (in the familiar sense that 'dog' does apply to Rover, among other canines, but does not apply to Puss)".

(3.3) According to intuition (a) expresses a truth, namely, the truth "that there is a set X of which each non-self-applicable plural noun phrase is a member".

(3.4) According to Boolos' proposal about how to interpret second order formulas, (a) should be read as

\[
\text{(b) There are some plural noun phrases that are all the plural noun phrases that do not...}
\]
apply to themselves.

(3.5) Consider the plural noun phrase *plural noun phrases that are all the plural noun phrases that do not apply to themselves*; "Does that plural noun phrase apply to itself or not? Is it one such plural noun phrase or not?". It is if and only if it isn't.

(3.6) So Boolos' proposal leads into paradox --into Russell’s paradox.

(4) The above argumentation 'suggests' that there may be no universally adequate form of semantics (I think this means, a form of semantics that can be used whatever the domain of objects the language is intended to talk about is).

Let's comment on Grim's remarks:

In regard to (1). It is not clear to me what Grim thinks the distinction between formal and informal semantics consists in. It seems that his distinction differentiates between what are commonly regarded as interpretations of first or second order languages (i.e. the standard theories of truth for first and second order languages) and the interpretation of a second order language appealing to Boolos' proposal. It seems that the talk of 'appealing to natural language plurals in order to interpret second order quantification' makes Grim think that the interpretation of a second order language is going to be 'informal' since it will be in natural language, which is 'informal'. Any other theory of how to interpret a (first or second order) language, though, is also
given in natural language. How else would we do it? Not all the expressive power and complexity of natural language is used in giving such an interpretation, though. Only a particularly suitable part (lacking ambiguities, etcetera) is used. But this is so for both what Grim regards as formal and as informal semantics. For instance, it seems that the following clause is part of what Grim regards as formal semantics

A sequence \( s \) satisfies \( \exists v G \) iff there is a set \( a \), and there is sequence \( t \) such that
t\( (v) = a \) and for any variable \( u \), if \( u \neq v \) then \( t(u) = s(u) \), and \( t \) satisfies \( G \)

whereas it seems that he would say that the following clause is part of an informal semantics

Some sets, \( \text{the } R\text{'s} \), and a sequence \( s \) satisfy \( \exists v G \) iff There are some sets, the \( X\text{'s} \), and there are some sets, \( \text{the } T\text{'s} \), such that: any given set \( x \) is one of the \( X\text{'s} \) if and only if the pairing of \( V \) and \( x \) is one of the \( T\text{'s} \), and for any second order variable \( U \), if \( U \neq V \) then for any given set \( x \) the pairing of \( U \) and \( x \) is one of the \( T\text{'s} \) if and only if it is one of the \( R\text{'s} \), and the \( T\text{'s} \) and \( s \) satisfy \( G \).

---

8 Could it be that with the talk of 'formal' and 'informal' semantics Grim is referring to a distinction like the one we draw in the third chapter between formal languages and regimented languages? I do not think so. Second order formal languages (in the sense introduced in section 1-(1) of chapter three, i.e. languages that we interpret by providing a theory of truth) are equally 'formal' or 'informal' under the standard truth theory or under Boolos' truth theory; and the same is true of second-order regimented languages under the standard interpretation (which involves translation in terms of classes) and under Boolos' interpretation (which involves translation in terms of plurals).
I do not see any essential difference between the two clauses. It might at most be argued that there is a difference of degree on how much intuitively clear natural language singular and plural quantification are. But Grim does not give any argument for this latter claim.

In regard to (2): What are the Russellian problems Grim is referring to? They have to be the paradoxes that Russell pointed out we would run into if there were a set $S$ of all sets: on the one hand we could ask for the cardinality of $P(S)$ which should be no greater than the one of $S$, since $P(S) \subseteq S$, but it should also be strictly greater than the cardinality of $S$ by Cantor's theorem. On the other hand if $S$ is a set then the elements of $S$ that are not members of themselves also constitute a set, let's call it $B$; but then $B$ is an element of $B$ if and only if it isn't. It seems that it is this second paradox (in any of its forms) that Grim has in mind when he talks of 'the Russellian problems'.

What is the relation of the 'Russellian problems' with the semantics of a second order language? The standard way of giving a semantics for a second order language involves interpreting second order variables as sets of the objects in the domain. In some cases, as with set theory, the objects in the domain may not constitute a set. This imposes a constraint on how to interpret second order sentences (either we allow second order variables to have as values not only sets but also proper classes, or we limit the domains second-order sentences may talk about to those domains which are sets). That the objects in the domain might no constitute a set is shown by the Russellian paradoxes. So the relation between the 'Russellian problems' and the semantics for second order languages is that the Russellian problems show something to be the
case (i.e. some objects might not be such that they are all and only the elements of some set), and this being the case imposes constraints on how to interpret second order variables.

Boolos' proposal tells us how we can express what we want to express by means of second order sentences if we interpret them adequately. It tells us how we can use second-order sentences to talk about some objects even if they do not constitute a set. So it overcomes a difficulty for other conceptions of how to interpret second-order languages. This difficulty for other conceptions stems from Russell paradox. And so we may say that Boolos' proposal is an attempt to 'escape from the Russellian problems'. But one should not get confused by this way of talking: Boolos' proposal does not say anything about sets or set theory, only how to interpret, among all the other monadic second-order languages, the second-order language of set theory; in particular, Boolos' proposal does not say anything at all about Russell's paradox if we take that paradox to consist in the observation that something which would seem true --given some sets, there is a set which has them as members-- leads to contradiction. It is as much a paradox after Boolos' proposal has been made as it was before.

In regard to (3): Before going into the main issue of (3), I would like to make two minor comments in regard to 3.1 and 3.3.

Grim's request in 3.1 to consider a "language rich enough to quantify over plural noun phrases and to express some of their basic properties" does not seem to be necessary: if we take the plural noun phrases as primitive objects in our domain (as it seems he does --in opposition
to take them as numbers or sets (coding the expressions)) then any first-order language is 'rich enough' to quantify over them; and as for the enough richness to express "their basic properties", he is only considering the relation 'applies to' which can be formalized by any two-place first order predicate. So instead of the sentence quoted above, it would have been more adequate just to say "consider a language with a two place predicate, which we intend to interpret as ...".

In 3.3, Grim claims that intuition tells us that the sentence

(a)  \( \exists X ( \exists x Xx \& \forall x [Xx \leftrightarrow \neg Axx] ) \)

expresses a truth. I would say that intuition can tell us something about the truth value of a sentence in a formal language only with respect to some interpretation of that language. Intuition by itself does not tell us that (a) expresses the truth that there is a set of plural noun phrases such that... ; intuition could only tell us that if we interpret the first-order variables as ranging over plural noun phrases, and "A" as the relation 'applies to', and interpret second order variables as having sets of the individuals in the domain as values, then the sentence is true (actually we will see that (a) is not true with respect to this interpretation, though).

Let's consider now the main argument in (3). "A" is interpreted as the relation 'applies to'. Regarding this relation Grim gives the following examples: "dog" applies to Rover, since Rover is a dog; "dog" does not apply to Puss, since Puss is not a dog; and to ask whether "plural noun phrases that are all the plural noun phrases that do not apply to themselves" applies to itself
is, according to Grim, to ask whether "plural noun phrases that are all the plural noun phrases that do not apply to themselves" is one such a plural noun phrase. So 'applies to' is a relation between a noun phrase and an object (which may or may not be a noun phrase). We might try to explicate this relation in the following way

For all x, for all a:

x applies to a if and only if x is a noun phrase and a is in the extension of x

Even if we have this explication we might think that the notion of 'applies to' is not clear enough if we want to use it in a general way (for instance, in the example above it does not seem that "plural noun phrases such that ..." applies to one object in the same way that a non-plural predicate like "dog" does). I believe that, without affecting what Grim wants to do with this notion of 'applies to', we can make it more precise if we focus on its use with respect to certain objects:

Grim's intention in appealing to the relation 'applies to' is to be able to arrive at a certain contradiction involving plural noun phrases. To get this we do not have to consider the relation in general --i.e. as holding or not for any two arbitrary objects. We can make it more precise (and still be able to arrive at the same kind of contradiction involving plurals) if we consider the relation restricted to a specific domain containing only certain noun phrases: noun phrases consisting of a plural common name possibly preceded by some adjectives and possibly followed by a relative clause. Then we can explicate this restricted sense of 'applies to' in the following
way:

If \( NP_i \) is a plural noun phrase and \( NP_j \) is a plural noun phrase, then

\[ NP_i \text{ applies to } NP_j \iff " " \sim NP_j \sim " " \sim " \text{is one of the} \sim NP_i \text{ is true} \]

Notice that whether we consider it in the general or the restricted sense, 'applies to' is a semantic relation. By this I mean that whether it holds or not of certain noun phrase and another object will not depend on the noun phrase as a syntactic object, but on what the noun phrase means. Whatever way we chose to explicate what the relation is we will have to appeal (as we did above) to notions like 'extension', 'truth', 'meaning' or 'denotation'.

According to the explication above of the restricted sense of 'applies to' we have that "dogs that bark at night" does not apply to "plural noun phrases with more than three words" since "plural noun phrases with more than three words" is one of the dogs that bark at night" is false; but that "plural noun phrases with more than three words" applies to "dogs that bark at night", since "dogs that bark at night" is one of the plural noun phrases with more than three words" is true.

And also we have that:

\[^9\text{I use the sign } "\sim" \text{ to express concatenation of expressions. So, for instance, } "\text{John}\sim\text{runs}" = "\text{Johnruns}".\]
if

"plural noun phrases that are all the plural noun phrases that do not apply to themselves" applies to itself

then (according to the explication above),

the sentence ""plural noun phrases that are all the plural noun phrases that do not apply to themselves" is one of the plural noun phrases that are all the plural noun phrases that do not apply to themselves" is true,

and so,

"plural noun phrases that are all the plural noun phrases that do not apply to themselves" is one of the plural noun phrases that are all the plural noun phrases that do not apply to themselves,

and so,

"plural noun phrases that are all the plural noun phrases that do not apply to themselves" does not apply to itself.

But, if

"plural noun phrases that are all the plural noun phrases that do not apply to themselves" does not apply to itself

then

the sentence ""plural noun phrases that are all the plural noun phrases that do not apply to themselves" is one of the plural noun phrases that are all the plural noun phrases that do not apply to themselves" is true,

and so (according to the explication above)

"plural noun phrases that are all the plural noun phrases that do not apply to themselves" applies to "plural noun phrases that are all the plural noun phrases that do not apply to themselves"

and so,

"plural noun phrases that are all the plural noun phrases that do not apply to themselves" applies to itself.

CONTRADICTION
So we arrive at a contradiction (as it is claimed in 3.5).

Now, where does this contradiction stem from? Grim’s answer (according to 3.6 and 4) is that it comes from quantifying over plural noun phrases while interpreting second order quantification in terms of plurals. I do not think this is right. If we consider, for instance, the explication above, the contradiction stems only from the explication of "applies to" (and the corresponding instance of the T-schema). I think that Grim thinks that the contradiction arises from quantifying over noun phrases while interpreting second order quantification in accordance with Boolos’ proposal because he believes that there is a parallelism between how the contradiction emerges in this case and how one can derive a contradiction by unrestrictedly interpreting the second-order quantifiers of the language of set-theory when appealing to sets. If we made it explicit, the parallelism that Grim seems to assume would be the one between (i)-(iii) and (i)'-(iii)’:

(i) If we interpret second-order quantification in terms of sets and, for instance, we take the sentence $\forall X \forall z \exists y \forall x (xy \leftrightarrow [x z \& Xx])$ to express the (unrestricted) separation principle, then we are assuming that there is a set of all sets.

(ii) So there is a set, let’s call it B, which has as elements the sets in the set of all sets which are not members themselves.

(iii) Then we can ask whether B is an element of itself, and it is if and only if it isn’t.

-120-
(i)' If we interpret second-order quantification in terms of plurals, then the interpretation of the true sentence (a) $\exists X(\exists xXx \& \forall x[Xx \leftrightarrow \neg Axx])$ will be the true claim that there are some plural noun phrases that are all the plural noun phrases that do not apply to themselves.

(ii)' So "plural noun phrases that are all the plural noun phrases that do not apply to themselves" is a plural noun phrase.

(iii)' Then we can ask whether that plural noun phrase applies to itself, and it does if and only if it doesn’t.

We need (i) and (ii) to arrive at the contradiction in (iii), but I think that for the plurals case we do not need (i)' and (ii)' to arrive at the contradiction. There is no need to justify that "plural noun phrases that are all the plural noun phrases that do not apply to themselves" is a plural noun phrase: this is a syntactic fact. I think that the only explanation for why Grim talks about the sentences (a) and (b), though, is because he thinks that there is the parallelism I tried to spell out explicitly above.

The contradiction, though, would also arise if we did not interpret the quantifiers in (a) in terms of plurals but in terms of sets: if there is a set whose members are all and only the non self-applicable plural noun phrases, then "plural noun phrases that are all the plural noun phrases that do not apply to themselves" is in that set if and only if it isn’t.
But we do not need to appeal to a second-order language to get the contradiction: if in a first-order language we have, for instance, a constant $c$ and we interpret it as the noun phrase "plural noun phrases that are all the plural noun phrases that do not apply to themselves", then the sentence $\neg Acc$ is true if and only if it isn’t.

Furthermore, we do not need to appeal to a formal language at all to get the contradiction, not even to plural noun phrases. We could get the same kind of contradiction with singular noun phrases: Does the singular noun phrase "singular noun phrase that does not apply to itself" apply to itself or not? It does if and only if it doesn’t.

Or instead of noun phrases we could be considering expressions in general ("expression that does not apply to itself" applies to itself if and only if it doesn’t), or adjectives (if an adjective is heterological if and only if it does not apply to itself, then "heterological" is heterological if and only if it isn’t).

So, in all these cases what we have is essentially Grelling’s heterological paradox. Grim’s criticism of Boolos’s proposal amounts to pointing out that the heterological paradox exists. But that should not be a criticism at all since Boolos makes no claim of having offered a ‘solution’ to it.

---

In our second explication of the relation 'applies to' it is clear that the contradiction arises only from how the relation 'applies to' works: certain noun phrase applies to another if and only if certain sentence is true, and this sentence may include those noun phrases and the relation "applies to", and so the door for contradiction is open.

Regarding (4): nothing of what has been said above prevents to apply Boolos' proposal about the interpretation of second order languages to a language where we quantify over plural noun phrases, whether we take them as primitive or as coded by numbers or sets. Notice that, unlike what Grim seems to suggest (p. 153), the difficulty of interpreting second order quantification appealing to sets does not depend essentially on the fact that we are quantifying over sets, i.e., it does not depend essentially on the fact that we are quantifying over the same kind of objects we use in the metalanguage to interpret quantification. If we were quantifying over, say, angels, and it so happened that there are many of them --for each set one angel-- then there would also be truths that we would not be able to express if second-order quantification is interpreted in terms of sets. (For instance "Given some angels, there are some angels who are all the angels who are not among the former ones" would not say the same as \( \forall X \exists Y \forall x (Yx \leftrightarrow XXx) \), if we interpret that formula as 'given some set of angels there is another set whose members are all the angels who are not members of the former set').
4 Michael Resnik's criticisms

Michael Resnik claims in his article *Second-order logic still wild*\(^{11}\) that Boolos' proposal of interpreting second-order quantification in terms of natural language plurals does not succeed in showing that second-order quantification does not carry a commitment to sets or classes of the objects the first-order variables range over. I think that his arguments can basically be divided into two parts. I will examine them in two sections.

4.1 Natural language plurals and commitment to classes

In the first part of his paper Resnik tries to make plausible that, in opposition to what Boolos claims, certain English sentences containing plurals do convey a commitment to classes.

(a) He starts by pointing out that, according to his intuitions, sentences like, for instance, "There are some critics such that any one of them admires another critic only if the latter is one of them distinct from the former" seem 'to refer to collections quite explicitly'. 'How else are we to understand the phrase "one of them" other than referring to some collection and as saying that the referent of "one" belongs to it?' (p. 77).

I think that Boolos himself contributes to giving these intuitions all the strength they can possibly have. The reason is this: Boolos (specially in TB) makes the reader consider a series of sentences some of which can be formalized in a first-order language, and some of which can only

be formalized in a second-order language; the reader is then aware of the existence of a
difference between these two kinds of sentences; the difference consists in being firstorderizable
or not; at this point --before Boolos' proposal has been presented-- the reader is understanding
second order quantification in the formalizations in the standard way, i.e., in terms of sets; so the
reader is led to understand the difference between the sentences that are firstorderizable and the
ones that are not as a difference between involving or not a reference to sets of the objects the
sentence is about; so the reader becomes convinced that certain sentences containing plural
quantification involve a commitment to a class of the objects the sentence is about.

(b) After stating his discrepant intuitions, Resnik considers Boolos' arguments that would show
that Resnik's intuitions are not the right ones. He considers the argument (in TB 447) that we
called ARG: the sentence "There are some sets that are self-identical, and every set that it is not
a member of itself is one of them" says something trivially true but if it involved reference to
a class would be false. First, Resnik questions whether, as Boolos claims, the sentence is really
true. Then he goes on into considering the argument Boolos gives to support that the sentence
is true. He considers the version of it in NP (pp. 331-332). As we mentioned in 1.3(c), Boolos
claims that it is intuitively clear that the sentences (1)-(3) say the same and, so, if there is no
quantification over classes in (3) there is none either in (1):

(1) There are some As such that every B is one of them
(2) There are some As and every B is one of them
(3) There are some As and every B is an A
Resnik claims that they do not say the same. They are naturally formalized as:

\[ (4) \exists X ( \exists x Xx \land \forall x [ Xx \rightarrow Ax ] \land \forall x [ Bx \rightarrow Xx ] ) \]

\[ (5) \exists X ( \exists x Xx \land \forall x [ Xx \rightarrow Ax ] \land \forall x [ Bx \rightarrow Ax ] ) \]

\[ (6) \exists x Ax \land \forall x ( Bx \rightarrow Ax ) \]

These sentences are equivalent within second-order logic. Nevertheless Resnik claims, and I agree with him, that we feel that, even if they are equivalent, there is a difference between (3) and the other two. (Actually, I would say that (2) is ambiguous: on the one hand it can be understood as (1), on the other it can be understood as (3)). Resnik's explanation of this difference is that they involve different ontological commitments. They are equivalent in the same way that 'John runs' and 'John belongs to the set of the runners' are; but like these two sentences, they also involve different ontological commitments. So, the equivalence holds only under certain ontological assumptions: in the last example, the existence of a set of all the runners, in (1)-(3) the existence of a set of As among which are all the Bs. Since the existence of an ontological commitment to classes in (1) would explain the contrast we feel there is between (1) and (3), the existence of these contrast gives some evidence for the quantification in (1) being a quantification over classes of the individuals in the domain of discourse.

So if Resnik's observations are correct he would have accomplished two things:

(i) showing that ARG fails, and
(ii) he would have given some positive evidence that sentences like (1) or (2) involve quantification over classes.

(c) Resnik gives (p. 80), still, another reason against Boolos’ argument that plural quantification does not involve commitment to classes: even if (contra what he has argued above) Boolos had succeed in showing that there is no quantification over classes in the sentence "There are some sets that are self-identical, and every set that is not a member of itself is one of them", he would not have shown that, in general, plural quantification does not convey commitment to classes: the sentence that Boolos considers is special since it is equivalent to a first-order one. It could be that that sentence did not involve commitment to classes but that the ones whose formalization does not have a first order equivalent did.

Let’s comment on Resnik’s arguments.

(d) Regarding the criticism of ARG: I think that noting that the sentence considered in ARG is intuitively true but that it would be false if it contained a reference to sets is a strong argument for the non-commitment to classes view. I believe, though, that the argument based on (1)-(3) that Boolos gives as additional support for ARG --to support the intuition that the sentence is intuitively true-- is much weaker than ARG itself.

Resnik does not offer any direct argument against ARG. All he does is criticize the 'supporting' argument based on (1)-(3), and regards doing so as having shown that ARG fails.
He has not shown that, though, since arguing against the justification based on (1)-(3) is not enough to show that ARG is not correct.

(e) What we just said in (d) shows that Resnik does not succeed in the first of the two points I mention at the end of (b) (i.e., in showing that ARG fails). In (f) and (g) we will examine the second point (i.e., that he has given positive evidence for sentences like (1) and (2) involving reference to classes). First, though, I want to present an augmented version of ARG which takes into account some possible ways in which the proponent of the commitment-to-collections view could try to escape the force of ARG:

The sentence

(7) There are some sets that are self-identical, and every set that is not a member of itself is one of them

is intuitively true but it would be false if it involved reference to sets. So plural quantification is not quantification over sets of the objects in the domain.

It seems that the one who does not want to accept this conclusion could try to reply in one of the three following ways:

(i) (the objection already explained in (c)) ARG involves a sentence which can be taken to be
firstorderizable (as in (6): \(\exists x \ x = x \& \forall x (x \in x \rightarrow x = x)\)). Even if the quantification in the sentence in ARG is not quantification over classes, plural quantification in other sentences might be.

We can reply to this by providing another sentence that could also be used in an argument like ARG and which is not firstorderizable: "There are some sets that are smaller than any set that is bigger than any of them, and every set with only one element is one of them".

(ii) (Resnik’s main reply) Even though at first our intuitions may tell us that (7) is true, if we think about it more carefully we realize that it is actually false, since it says that there is a set of all sets.

To this it can be replied: consider the sentence "The sets that are not members of themselves do not constitute a set"; it seems that our intuitions about the truth of this sentence are even stronger than for (7) (Resnik himself might have uttered it sometime); so, "The sets that are not members of themselves" can not mean 'the set of sets that are not members of themselves'.

The Resnikian might reply: your argument only shows that some referential plural noun phrases with collective predicates do not involve reference to sets, but this does not show that plural quantification in expressions of the type "there are some As that..." does not involve quantification over sets. To this I would reply: the sentence "The sets that are not members of themselves do not constitute a set" implies "there are some sets that are all the sets that are not..."
members of themselves" (if the latter is false the former would either be false or lack meaning); so, since the former is true the latter is also true, and so, plural quantification is not quantification over sets of the objects in the domain.

(iii) The third possible reply to ARG would be to accept proper classes (collections of sets that are not themselves sets, and so that are not themselves members of proper classes) and to claim that plural quantification corresponds to forming an unity or collection out of whatever one is quantifying over. So we do not have to deny our clear intuitions regarding the truth of "There are some sets that are self-identical, and every set that is not a member of itself is one of them". The sentence is true since it is true that there is a (proper) class that contains all sets.

Boolos' idea can also be applied to this case, though\textsuperscript{12}: it seems intuitively true that there are some things which are all the sets and proper classes that are not members of themselves. But it is false that there is a class of all the sets and proper classes. So plural quantification is not quantification over sets or proper classes of elements in the domain.

To this one might reply in two ways, analogous to (ii) and (iii)\textsuperscript{13}: [first reply] (ii)' the claim in the previous paragraph (there are some things which are all the sets and proper classes that

\textsuperscript{12} I take the idea of applying Boolos' argument to proper classes, 'awesome classes', etc. from David Lewis, \textit{Parts of Classes} (1991), Basil Blackwell, pp. 67-68.

\textsuperscript{13} It does not seem possible to reply in a way analogous to (i) (i.e. to say that the claim that there are some sets and proper classes that are \textit{all} the sets and proper classes that are not members of themselves is formalizable in first-order as in (6)), since a formalization with the form of (6) would not 'capture' the sense conveyed by the word "all".
are not members of themselves) is actually false; To this, a contra-argument analogous to the one in (ii) would apply. [second reply] (iii)’ there are 'awesome collections’ which contain even the proper classes as members, and so the sentence above means that there is an awesome collection which contains all and only the sets and proper classes that are not members of themselves which is something true. Counter-argument: the idea in ARG also applies to awesome collections: if there are awesome collections, then it seems true that there are some things that are all the sets, proper classes and awesome collections that are not members of themselves, and so this claim is not the same as claiming that there is a set, proper class or awesome collection whose members are all the sets, proper classes and awesome collections that are not self-members, since in this case the claim would be false.

For any ultimate kind of making-one-out-of-many objects that a singularist might postulate, the type of argument in this section will apply to these objects and show that plural quantification can not involve that kind of collecting over the objects we are talking about.

There is still a position that the singularist may hold: there is no ultimate kind of making-one-out-of-many objects. So it might seem that he can always escape the argument in this section by going 'one step higher'. But this is not quite so, since we can consider the intuitively true sentence "there are some things such that every thing is one of them" or "there are some things which are all what there is". This can not mean that there is a collection-like entity that contains everything, since then the sentence should be false.
It might be replied that the sentence is true but that it does not say what we might have thought: the sentence says that there are some things that are all the things we are talking about, but this does not mean that they are 'absolutely everything that existst'. This is so because it is not possible to quantify over absolutely everything. One might be tempted to reply to this that the one who says that it is not possible to quantify over absolutely everything is himself, when making this claim, actually quantifying over everything. The one who denies the possibility of completely general quantification could reply, though, that when saying it is not possible to quantify over everything he is just claiming that whatever our domain on certain particular use of quantification is, there is another possible domain of quantification that includes it. Still, if this more careful formulation really says what the foe of the possibility of quantifying over everything wants it to say, then it should be saying something about all domains of quantification; and if a domain is just the individuals we quantify over on some specific use of language, quantifying over all domains involves quantifying over all individuals.

Moreover, it seems quite obvious that if I say, for instance, that everything is identical to itself I am talking about everything that there is (there is no one object whose being or not being identical to itself is irrelevant for the truth of my statement). We should need good reasons for abandoning the view that we can quantify over everything. But as already argued by Richard

---

14 Objecting to the one who claims that we can not quantify over absolutely everything by saying that in making her claim she is already quantifying over everything, seems to be analogous to objecting to the one that claims that Santa does not exist by saying that in making her claim she is already committing herself to the existence of Santa.

-132-
Cartwright in his *Speaking of Everything* there simply does not seem to be any such good reasons\(^{15}\).

Notice also that if one believes that plural quantification is quantification over classes, and also believes that it is not possible to quantify over absolutely everything, then he can say that the intuitively true sentence "there are some things which are everything there is" is true, only at the price of having to admit classes that are members of themselves: if "there are some things" is just equivalent to "there is a class of things", and both "there is" and "every" are interpreted with respect to the same domain, then the class of objects must be one of the objects in the domain.

(f) Let's now comment on Resnik's criticism of Boolos' supporting argument for ARG based on the intuitive equivalence of sentences (1)-(3). As we saw, this criticism would not only show that the supporting argument for ARG has no force, but would also seem to provide some positive evidence for the non-commitment to classes view. As I mentioned before, I think that Resnik is right in saying that there is a difference in meaning between (3) and the other two sentences. Consider, for instance, the following sentences corresponding to (1) and (3) respectively:

(8) There are some shops on my street such that every pet-shop in town is one of them

(9) There are some shops on my street and every pet-shop in town is a shop on my street

The difference between them becomes clearer if we add another sentence to them:

(10) There are some shops on my street such that every pet-shop in town is one of them. They are the mayor’s favourite shops.

(11) There are some shops on my street and every pet-shop in town is a shop on my street. They are the mayor’s favourite shops.

(10) is a correct text and "They" in the second part refers to the shops the first part talks about --some shops which are in my street and among which are all the pet-shops in town. But (11), unlike (10), is strange if not unacceptable, since it is not clear what "they" refers to.

Someone might claim that the difference pointed out above is only a syntactic/grammatical one with not clear bearing on questions of meaning and ontological commitment which is what we are interested in here. I do not think that this is correct: both clauses in each text are grammatical, and they are independent at the syntactic level; the relation between the two is only at the level of their meaning. (I think that what I say below will make this clearer).

So there is a difference between the two sentences. Contra Resnik, though, I do not think that this difference has to do with different ontological commitments. We discuss this question in sub-section (g).
The use of natural language always takes place with respect to a background of common beliefs and assumptions—a so called context. The context in which natural language use takes place provides the domain of individuals we are talking about. So, for instance, we say that in the sentence "The math teacher is sleeping", the expression "the math teacher" denotes that unique individual that is a math teacher, but we do not mean, of course, 'the only being in the universe who teaches math', but rather 'the only person in our domain of discourse—in our context-- who teaches math'. Also the context is crucial in determining the referent for such expressions as "I", "tomorrow", "here", etc.

Sometimes it is said that the idea of interpreting formal language sentences by means of an interpretation that has certain domain corresponds to the idea that natural language expressions are interpreted against the backdrop of a context. I think that this is only partially true: the context in a conversation changes constantly, and so does the domain of objects we are talking about; but I do not think it is adequate to see this as a constant reinterpretation of the language using interpretations with different domains.\(^\text{16}\)

The context may contain one or several individuals who are 'salient' or 'prominent'. Expressions like "she" or "he", for instance, denote the only prominent female human or male human, respectively, in the context (if there is no such prominent individual then the use of the expression would be inadequate). Also the expression "they" will denote some individuals which are 'highlighted', 'salient' or 'prominent' in the context. In natural language once an individual

\(^{16}\) For more on this idea see section 2 in chapter 3.
has been introduced or made prominent we can keep on referring to it by the use of pronouns. Analogously, once several individuals have been made salient we can keep on referring to them by the pronouns "they" or "them").

Two sentences might say the same (have the same truth conditions) but affect the context in a different way. This is what happens with (1)-(3). By uttering (1) ("There are some As such that every B is one of them") we affect the context in that it comes to have some prominent individuals --they are As and all the Bs are among them--. By uttering (3) ("There are some As and every B is an A") one describes a situation identical with the one described by (1) (there has to be some individuals which are As, and all the Bs have to be also As), but there are no individuals in the context made salient as a consequence of uttering the sentence. This explains the difference we observed between (8) and (9). (8) and (9) are equivalent but modify the context in a different way: (8) introduces some individuals as prominent (some shops in my street among which are all pet-shops in town) whereas (9) does not. This is why after uttering (8) we can add "they are the mayor’s favourite shops" since "they" will refer to that salient individuals, but if we add it after (9) the text is non-felicitous, since there is no clear referent for "they"\(^{17}\).

\(^{17}\) Another example illustrating the same point which involves just one jingle individual: Consider the sentences

(a) There is exactly one typo in this footnote.
(b) It is not the case that there are no typos in this footnote and it is not the case that there are two or more typos in this footnote.

They say the same (describe the same situation), but affect the context differently. This is why (c) is correct but (d) is bad

(c) There is exactly one typo in this footnote. It is in the first line.

-136-
This diverse way in which (1) and (3) affect the context is also what explains that even if we realize that they are equivalent, we feel that there is a difference between them. This difference, though, has to do with the different way the two sentences affect the context where they are used, and not with a distinction in the ontological commitments of each sentence. So, contra Resnik, we do not need to believe that (1)-(3) convey different ontological commitments in order to explain the difference we feel there exist between them.

4.2 Second-order quantification and natural language plurals

(a) According to Resnik, the arguments he gives that we explained in the previous section have established that Boolos does not succeed in showing that natural language sentences containing plural quantification do not involve commitment to classes. Then he goes on into considering what is the main issue on Boolos’ proposal: whether we can interpret second-order languages’ quantification without appealing to classes.

He considers Boolos’ theories of truth for the first and second-order set theory languages in NP. He claims that the use of sequences in those definitions obscures what the is commitments are. For that reason he wants to offer a truth theory of the same kind as Boolos’ but where the commitment of existential quantifiers is displayed more perspicuously.

(d) It is not the case that there are no typos in this footnote and it is not the case that there are two or more typos in this footnote. It is in the first line.
He recursively defines truth for an interpretation $I$ in a domain $D$. In the first order case an interpretation relates each variable of the language with a unique element of $D$; in addition to that, in the second-order case the interpretation also relates each second order variable with $0$ or more elements of $D$.

The clause for the existential quantifier in the second-order case is:

$$\exists VS \text{ is } T \text{ for } I \text{ in } D \text{ iff there is a relation } R \text{ over } D \text{ consisting solely of } 0 \text{ or more ordered pairs of the form } <V,d> \text{ (} d \text{ in } D) \text{ such that } S \text{ is } T \text{ for } I[V/R] \text{ in } D$$

(where $I[V/R]$ is like $I$ except for containing the pairs in $R$ in place of the pairs $<V,d>$)

If we interpret "there is a relation $R$" as 'there is a class of ordered pairs' then it is clear than we are committing ourselves to classes; if we interpret it as 'there are some ordered pairs', then the existence of a commitment to classes will depend on whether the plural quantification conveys such a commitment or not.

So whether it has been shown that there is a way of giving a theory of truth for the second-order language of set-theory without interpreting the variables in terms of classes or not will depend on whether the plural quantification "there are some ordered pairs" conveys a commitment to classes. He claims that 'perhaps the most straightforward method' to find out whether it does or not is 'to see it as a question of deciding whether such sentences should be represented in second-order logic or in class theory' (p. 83). He says thatBoolos himself seems
to endorse this suggestion, but that it would be 'of no avail unless we already know what ontic commitment second-order quantification carries'. But what the commitment of second-order quantification is is precisely what we were trying to find out in the first place. So, Resnik concludes, 'Boolos is involved in a circle'.

To this I think that the following could be replied.

As we pointed out, Resnik thinks that Boolos has not succeeded in making it plausible that plural quantification is not quantification over sets of the objects in the domain. Even if Resnik were right on this, the charge of circularity would not be justified, since it is not true that Boolos endorses that to find out the ontological commitment of a sentence containing plurals we should see whether it should be formalized in second order-logic. Resnik does not mention in which of the articles and in which place in the article Boolos makes such a claim; neither does he say which of the claims that Boolos makes would justify inferring that Boolos' endorses such a view.

When Boolos argues that natural language plural quantification does not convey commitment to classes he does not appeal to second-order quantification. There is no circle: first (first at the logical level, not necessarily at the expository level) he makes plausible that plural quantification does not imply commitment to classes, then he gives a theory of truth for a second-order language using natural language plurals. If plurals do not involve a commitment to classes, then there is no appeal to classes in the truth theory. And that plurals do not convey
commitment to classes has previously been made plausible. If plural quantification were quantification over classes then his proposal would fail, but still there would be no circle in his argumentation.

(b) The last argument in Resnik's paper that we will examine also deals with the relation between plurals and second-order quantification (it is in p. 87): Let's assume that Boolos is right in claiming that sentences like (1)-(3) are synonymous. Then we have that there are two kinds of sentences containing plural quantification: the 'irreducibly plural' which can not be formalized by means of a first-order sentence, and the 'non-irreducibly plural' which are firstorderizable. What does differentiate the two kinds of sentences? Resnik suggests that a way of explaining the difference would be to point out that there are two senses of "There are some things such that" and that in one sense it just means 'There is at least one thing such that', and that in the other it means 'there is a collection of things such that'. But, of course, this is something Boolos would not want to say. So, Boolos does not have a way of explaining what the difference between the two senses is. Resnik says that 'that puts him in the ironic position of saying that in the other sense it simply means what it says'.

It seems to me that to make his point about the difficulty in explaining what the two ways of understanding sentences with plural quantification consist in Resnik did not have to assume that (1)-(3) are synonymous. He could have pointed out to the difference we feel there is between the sentences which are most naturally formalized in second-order and the ones which are not, and ask what the difference between the two consist in. Whatever the way the question is
presented, though, I think that the following could be replied:

There are not two senses of "There are some things such that". Both in the sentences which are formalizable in first-order and in the ones which are not it means the same: it selects some arbitrary individuals of which something is going to be predicated. The difference between the two kinds of sentences lies in the predicate. There is a relation between firstorderizable sentences and sentences with distributive predicates, and also between non-firstorderizable sentences and sentences with certain collective predicates. When the predicate is distributive we will be considering the application of the predicate to each of the arbitrary things selected by "There are some things such that"; if the predicate is collective then it applies only with respect to all of them. There are certain constructions such as "each other" and "one of them" that can be used in forming a complex collective predicate out of a (possible non-collective) simple predicate; for instance 'are such that Daniel painted one of them' is collective even though 'were painted by Daniel' is distributive; or also, 'are faster than Zev and also faster than the sire of any horse that is slower than all of them' is collective, even though 'are faster than x' is distributive.

Now we can explain the difference that Resnik suggests can only be explained by appealing to classes, without appealing to them: The sentences containing plural quantification that are firstorderizable are the ones for which there is one equivalent sentence whose predicate is distributive; the sentences that are non firstorderizable are the ones for which there is no such equivalent sentence.
If instead of considering the difference 'firstorderizable/non-firstorderizable' we consider, as I suggested, the difference 'most naturally formalizable in first-order/most naturally formalizable in second-order' we have that: the sentences whose most natural formalization is in second-order are the ones containing collective predicates constructed on the basis of a distributive predicate; the sentences whose most natural formalization is in first-order contain distributive predicates.
Chapter Three

On the Interpretation of Formal Languages and the Analysis of Logical Properties

We will be examining some aspects of the relationship between natural language and the so called formal languages. We can not pretend to be making claims about all formal languages, since there are infinitely many different kinds of sign systems that could be regarded as formal languages, and we would not know even how to approach the task of trying to say anything about all of them. We will restrict our attention to the languages of standard propositional logic and standard first and second order logic. Almost all of the time we will focus our attention specifically to standard first order languages.

We will approximate the issue of the relationship between natural language and its formalizations by means of formal languages by considering first the following question: What does a sentence of a formal language mean?

1 We will be making three distinct uses of italics: (i) for emphasis; (ii) to talk about an expression type of which we are exhibiting a token, i.e. we might use italics instead of quotation marks; (iii) to talk about expressions type while exhibiting tokens of some of them and/or using metavariables, i.e. we might use italics instead of corners. We do this for simplicity. Only when being precise becomes essential we will have recourse to quotation marks or corners.
1. Formal Languages and What They Mean

I think we can distinguish at least four main senses in which we say that a sentence of a formal language, and specifically, of a first order language, has certain meaning. For our purposes the two important senses will be the ones we will consider in subsections (1) and (3) below. The four senses are the following:

(1) We could say that a sentence of a formal language does not by itself mean anything unless we interpret it, and to interpret it consists in providing in the standard way a so called model for it (models are also called interpretations, or structures).

There are different ways of specifying what a model for a first order language L is. One common way of doing it is to say that a model M is an ordered pair <D,F> such that D is a set, the so called domain (or universe) of M, and F is a function that assigns an appropriate value to each non logical primitive symbol of L: an element of M to each constant, a subset of D to each 1-place predicate symbol, a subset of n-tuples of elements of D to each n-place predicate symbol (n≥2), and a subset of n+1 tuples to each n-place function symbol.

A model by itself does not yet endow the formal language L with meaning. If it does so, it is only with respect to a theory that tells us what the interpretation or the value of complex expressions is, and specifically, what the interpretation of the sentences is. There are some differences in the specific form that such a theory can have. We will consider here two slightly
different presentations which are both standard. (Several other presentations are possible, including some which are hybrids of the two considered here).

One way to proceed is to provide a truth theory for L and to do this through a definition of satisfaction: we define first that a model M (=<D,F>) and an appropriate sequence (or assignment) s satisfy a formula $\alpha$ of L. (A sequence is a function whose domain is the set of variables, an appropriate sequence for M is a sequence whose range is D). Then we can say that a formula $\alpha$ is true in a model M if there is an appropriate sequence s such that M and s satisfy $\alpha$ (or alternatively, if for any appropriate sequence s, M and s satisfy $\alpha$). In order to define satisfaction for a first-order language it is common to proceed in the following way:

First, we give a recursive definition of the denotation (designation, or value) of a term t with respect to a model M and a sequence s, which we will notate as $M/s(t)$:

if t is a variable then $M/s(t)=s(t)$,

if t is a constant then $M/s(t)=F(t)$,

if t is $ft_1...t_n$ then $M/s(t)=F(f)(M/s(t_1),...,M/s(t_n))$

Then we give a recursive definition of $M$ and s satisfy formula $\alpha$ that has the following form (where M=<D,F>):

If $\alpha=t_1=t_2$, where $t_1$ and $t_2$ are terms, then $M$ and s satisfy $\alpha$ iff $M/s(t_1)=M/s(t_2)$.
if $\alpha = \phi t$, where $\phi$ is a monadic predicate symbol and $t$ is a term, then $M$ and $s$ satisfy $\alpha$ iff 
\[ M/s(t) \in F(\phi), \]
if $\alpha = R t_1 \ldots t_n$, where $R$ is an $n$-adic ($n \geq 2$) predicate symbol and $t_1 \ldots t_n$ are terms, then $M$ and $s$ satisfy $\alpha$ iff 
\[ <M/s(t_1), \ldots, M/s(t_n)> \in F(R), \]
if $\alpha = \neg \beta$, where $\beta$ is a formula, then $M$ and $s$ satisfy $\alpha$ iff $M$ and $s$ do not satisfy $\beta$,
if $\alpha = (\beta \land \gamma)$, where $\beta$ and $\gamma$ are formulas, $M$ and $s$ satisfy $\alpha$ iff $M$ and $s$ satisfy both $\beta$ and $\gamma$
if $\alpha = \exists x \alpha$, where $\alpha$ is a formula, then $M$ and $s$ satisfy $\alpha$ iff there is an element of $D$, $a$, such that 
$M$ and $s^a_x$ satisfy $\alpha$, where $s^a_x$ is a sequence that assigns $a$ to $x$ and which otherwise is just like $s$.

An alternative way to proceed in order to provide an interpretation for the sentences of the first-order formal language $L$ is the following: given a model $M = <D, F>$ for $L$, we recursively define the function $I$, which we might call the interpretation under $M$, that assigns a value to every primitive non logical expression of $L$, to every closed term and to every sentence of $L$. To each sentence of $L$ it assigns either the value True or the value False.

If $e$ is a constant, a function symbol, or a predicate symbol of $L$, $I(e) = F(e)$,
if $f$ is a $n$-place function symbol and $t_1, \ldots, t_n$ are closed terms then $I(ft_1 \ldots t_n) = I(f)(<I(t_1), \ldots, I(t_n)>)$,
if $\alpha = t_1 = t_2$, where $t_1$ and $t_2$ are terms, $I(\alpha) = \text{True}$ if $I(t_1) = I(t_2)$, and $I(\alpha) = \text{False}$ if $I(t_1) \neq I(t_2)$,
if $\alpha = \phi t$, where $\phi$ is a 1-place predicate symbol and $t$ is a closed term, then $I(\alpha) = \text{True}$ if $I(t) \in I(\phi)$, and $I(\alpha) = \text{False}$ if $I(t) \notin I(\phi)$,
if \( \alpha = R_{t_1, \ldots, t_n} \), where \( R \) is an \( n \)-place \( (n \geq 2) \) predicate symbol, and \( t_1, \ldots, t_n \) are closed terms, then

\[
I(\alpha) = \text{True if } <I(t_1), \ldots, I(t_n)> \in I(R), \quad \text{and } I(\alpha) = \text{False if } <I(t_1), \ldots, I(t_n)> \notin I(R),
\]

if \( \alpha = \exists x \beta \), where \( \beta \) is a formula, \( I(\alpha) = \text{True if } I_a^*(\beta_{xa}) = \text{True for some } e \in D, \quad I(\alpha) = \text{False if } I_a^*(\beta_{xa}) = \text{False for all } e \in \text{the domain of } D, \) where \( \beta_{xa} \) is a sentence obtained from \( \beta \) by replacing all free occurrences of \( x \) with a new constant \( a \) which does not appear in \( \beta \), and \( I_a^* \) is a function that assigns \( e \) to \( a \) and which otherwise is just like \( I \).

Now, let’s consider some specific formal language, say the language \( \mathcal{L} \) that has one constant symbol \( a \) and one predicate symbol \( P \), and some specific interpretation for the language, i.e. one model for the language, say the model \( M \) whose domain is the set of humans, that assigns David Armstrong to \( a \), and assigns the set \( \{x: x \text{ philosophizes}\} \) to \( P \).

Given this interpretation, does the sentence of \( \mathcal{L} \) \( Pa \) mean the same as the English sentence Armstrong philosophizes? I think it is clear that it does not. If we consider the second presentation given above we see that the 'value' or 'interpretation' that we assign to a sentence is either True or False. In our specific example, we would have that the value of \( Pa \) is True. All the other true sentences of \( \mathcal{L} \) would be assigned the same value as \( Pa \) by the interpretation function under \( M \). If what determines the interpretation of \( Pa \) in \( M \), i.e. what determines the meaning of \( Pa \) according to \( M \), is the value that the sentence gets assigned by \( I \), the interpretation function under \( M \), then certainly \( Pa \) does not mean the same as Armstrong philosophizes, or otherwise we would be equally justified in claiming that, for instance, \( \exists x Px \) means that Armstrong philosophizes, since \( \exists x Px \) gets assigned by \( I \) the same value as \( Pa \).
Maybe someone might argue in the following way: it is incorrect to take 'having the value True' as being all that the second approach above says about the interpretation of $Pa$. Given the way the interpretation function is defined it also tells us 'when' $Pa$ is true, namely when Armstrong philosophizes. So we would have that given the interpretation function or, at least, given the way it is defined, we can conclude that $Pa$ means that Armstrong philosophizes. This is obscured by the very fact that we use a function and we assign an object to each expression. We should understand the claim that $I(\alpha)=True$ as just another way of expressing that $\alpha$ is true. Viewing things this way the second approach is just like the first in that it is a way of providing a theory of truth for the language on the basis of a model.

I think that the view expressed in the previous paragraph is not correct. First, the claim that the value of $I(\alpha)$ is True, taken by itself, is a completely different claim from the claim that $\alpha$ is true. True is an object (or so me must assume if the definition of $I$ is to make sense) --an abstract one. So is $\alpha$. Given any two objects we can always define a function that will make one the value of the other, but this fact by itself does not imply anything about the two objects or their relationship other than we have stipulated that the function we have defined assigns one to the other. We could define another function $G$ that made blueness (if such entity exists) the value for the argument the flag of the People’s Republic of China. Then it would be the case that $G(China’s\ flag)=Blue$, but this does not mean that it would also be the case that China’s flag is blue. Analogously, the claim that $I(Pa)=True$ is a claim about which objects happen to be related by $I$, not about whether $\alpha$ is or not true. If we want the second approach to yield a theory of truth we should incorporate a clause such as: if $I(\alpha)=True$ then $\alpha$ is true, if $I(\alpha)=False$ then $\alpha$ is false.
Notice that given that we need a clause such as the one just stated, instead of postulating the range of $I$ to be the set \{True, False\}, we could postulate it to be the set \{1,0\} and then have the clause: if $I(\alpha)=1$ then $\alpha$ is true, if $I(\alpha)=0$ then $\alpha$ is false. The only difference between having one or the other set as the range for $I$ is that in the first case is easier to infer on the basis of $I$ (and the fact that $I$ is presented as an interpretation function) the clause that would allow us to obtain a truth theory.

Second, even if we left the considerations in the previous paragraph aside and considered the second approach basically as the same as the first one, i.e., as a way of providing a truth theory, it would still not be the case that interpreting $Pa$ in accordance with this second approach would make $Pa$ to mean that Armstrong philosophizes. This is so because the first approach does not make it the case either. Let’s see why it does not:

It is true that, given the model M above, the truth theory that the first approach provides would yield the following biconditional:

(a) $Pa$ is true iff Armstrong $\in \{x: x$ philosophizes\}$

But even if the truth theory yields this biconditional, it does not make it the case that $Pa$ means that Armstrong philosophizes. We can point two three sort of facts that show that this is so, the most relevant being the first one:
(i) The biconditional in (a) involves only material implication. That is, in order for the 
biconditional to be true all that is required is that the sentences appearing on the right and on the 
left of \( \text{iff} \) be both true or both false. Given that the set \( \{x:x \text{ philosophizes}\} \) does in fact have 
Armstrong as a member, (a) allows us to conclude that \( Pa \) is true. We could have obtained this 
extact same information if instead of (a) we had (a)'

\[
\text{(a)}' \ Pa \text{ is true iff Lennon was born in Liverpool}
\]

The sentence appearing in the right hand side of \( \text{iff} \) does not tell us 'when' \( Pa \) is true, only that 
it is true. From (a)' we can not conclude for instance: if Lennon had not been born in Liverpool, 
\( Pa \) would not be true.

Maybe it could be replied that what makes \( Pa \) mean that Armstrong philosophizes is not 
just that the biconditional (a) follows from the truth theory, but the whole interpretation for the 
language, including the interpretation of the expressions in \( Pa \). I do not think this is correct, 
though. Suppose we interpret the language \( \mathfrak{L} \) with respect to the same model as before, and with 
the following minor modification to the definition of satisfaction: instead of having the clause 
in (b) as before, we have (b)'

\[
\text{(b)} \quad \text{if } \alpha = \rho t, \text{ where } \rho \text{ is a monadic predicate symbol and } t \text{ is a term, then } M \text{ and } s 
\text{satisfy } \alpha \text{ iff } M/s(t) \in F(\rho)
\]
(b') if $\alpha = \varrho t$, where $\varrho$ is a monadic predicate symbol and $t$ is a term and $\varrho t \neq Pa$, then

$M$ and $s$ satisfy $\alpha$ iff $M/s(t) \in F(\varrho)$, if $\varrho t = Pa$ then $M$ and $s$ satisfy $\alpha$ iff

Lennon $\in \{x: x$ was born in Liverpool.\}

Every primitive symbol of $\mathcal{L}$ would still be assigned the same value as before. And all the sentences of $\mathcal{L}$ would have the exact same truth value. So, there seems to be no reason to claim that under one presentation of the interpretation of $\mathcal{L}$ Pa means that Armstrong philosophizes, but under the other it means something else -- we must keep in mind that a biconditional that follows from the theory does not say anything about any connection between what the two sentences on each side of the iff express; otherwise put: from the truth of a biconditional and what one of the two sentences expresses, we can not conclude anything about what the other sentences expresses, other than it expresses something that determines the same truth value as the one determined by what the former sentence expresses.

(ii) A second way of realizing that the fact that the truth theory yields (a) does not make it the case that $Pa$ means that Armstrong philosophizes is by noticing that sets are extensional. The set $\{x: x$ philosophizes\} is presumably the same as $\{x: x$ philosophizes and $x$ is a rational being\}, or as $\{x: x$ philosophizes and $x$ is not a new born\}. We would still have the same model $M$ if we had specified the value of $P$ as being the set $\{x: x$ is not a new born and $x$ philosophizes\}. We would then say that the truth theory would have as a consequence (c):

(c) $Pa$ is true iff Armstrong $\in \{x: x$ is not a new born and $x$ philosophizes\}
If having (a) as a consequence made it the case that under the interpretation induced by M \( Pa \) meant that Armstrong philosophizes, then if the theory yields (c) we would have to say that \( Pa \) means that Armstrong philosophizes and is not a new born. This is absurd since, as we pointed out, the model is the same no matter how we specify the set that is the value of \( P \).

(iii) A third way of realizing that the fact that the interpretation induced by M yields the biconditional (a) does not suffice to make it the case that \( Pa \) means the same as the English sentence *Armstrong philosophizes* is by noting that the English sentence does not say anything about sets or the membership relation, whereas the sentence in the right hand side of (a) is about the membership of an object in a set\(^2\). And even if we think that (a) is not by itself what determines what the meaning of \( Pa \) is, we should note that the value assigned to \( P \) is a set, and whatever we might want to say about how the value of \( P \) contributes to what \( Pa \) means, it is this set and no something else that will play a role.

There seems to be very good reasons, then, for thinking that \( Pa \) when interpreted in the standard way on the basis of M does not mean the same as the English sentence *Armstrong philosophizes*.

\(^2\) The position according to which asserting some predication is the same as asserting some membership relation does not agree with intuition and the burden of justification is on the side of the one that wants to hold that view. Still here are two reasons, in addition to its conflict with basic intuitions, for not holding it: (a) someone can believe that Armstrong philosophizes but not believe that Armstrong\( \in \{ x : x \text{ philosophizes} \} \); (b) if we accept the principle that assertion of predication is assertion of certain membership relation then we have to accept that the principle also applies to the assertion that Armstrong\( \in \{ x : x \text{ philosophizes} \} \), and so that this assertion is the same as: \( <\text{Armstrong}, \{ x : x \text{ philosophizes} \}> \in \{ <x,y> : x \in D \& y \in P(D) \& x \in y \} \); and the principle would also apply to this latter predication of membership, etc.; it becomes intuitively less and less plausible that all these other assertions of membership are the same as the original assertion that Armstrong philosophizes.
philosophizes. Does \( Pa \) so interpreted mean the same as any English sentence? Well, which English sentence could it be? It seems that the most plausible candidate would be *Armstrong is a member of \( \{x : x \text{ philosophizes}\} \).* It seems clear, though, that this English sentence will not do either. First, there is the fact that, as noted in (ii) above, whatever \( Pa \) might mean is not sensible to the different ways of specifying the set \( \{x : x \text{ philosophizes}\} \), whereas this is not true of the English sentence under consideration. Moreover: even if we were interested only in a notion of 'sameness of meaning' according to which the sentences \( \{x : x \text{ is author of Carrie}\} \text{ has one member} \) and \( \{\text{Stephen King} \text{ has one member}\} \) would have the same meaning, the sort of difficulty raised in (i) above would also apply to the candidate English sentence we are now considering: Given the same model \( M \) we can provide an alternative formulation of the truth theory that provides the same interpretation for each expression in the language, but which does not give any condition involving reference to any set when specifying the truth condition of \( Pa \) (we can use, for instance, a clause such as (b)').

It seems, then, that \( Pa \) does not mean the same as any English sentence. I believe that this fact makes it plausible to think that it does not mean anything at all, if for a sentence to mean something requires not just that it possesses some semantic property or other like, for instance, to include some expression that refers to some specific individual, but also that the sentence does 'the same sort of thing' that natural language sentences do. If \( Pa \) had the sort of meaning that natural language sentences have then it seems that anyone who understands what we do when we provide the interpretation of \( \xi \) on the basis of model \( M \), would understand what the meaning that \( Pa \) is endowed with is, and should be able to express it in English.
If $Pa$, when evaluated with respect to the model for $M$, does not mean anything, what do we do when we provide in the standard way a so called *interpretation* for a first order formal language? Do the expressions of the language have any sort of semantic property? We will try to say something about this later on, in section 3.

(2) Sometimes we might claim that, for instance, sentence (d) says that $R$ is transitive; or that sentence (e) says that there are infinitely many things; or that (f) says that nothing is $P$

(d) $\forall x \forall y \forall z (Rxy \land Ryz \rightarrow Rxz)$

(e) $\exists X (\forall x \exists y \forall z (Xxz \leftrightarrow z=y) \land \forall x \forall y (\exists z (Xxz \land Xyz) \rightarrow x=y) \land \exists x \forall y \neg Xyx)$

(f) $\neg \exists x P x$

These claims exemplify another sense of what a sentence of a formal language means. The claims here about what a sentence $\alpha$ means have to be understood as claims about what will be the case in all and only the models in which $\alpha$ is true. Furthermore, what we pretend to be claiming about some primitive symbol of $S$ appearing in $\alpha$ (‘$R$ is transitive’), is actually what will be true of the interpretation of that primitive symbol in each model where $\alpha$ is true. So, for instance, we say that (d) means that $R$ is transitive because (d) is true in all and only the models where the interpretation of $R$ is a transitive relation; or we say that (e) means that there are infinitely many objects because in each model in which (e) is true the domain will be an infinite set and, furthermore, (e) is true in all models with an infinite domain.
Maybe there is also a looser use of this sense of 'the meaning of $\alpha$' where a sentence of a formal language is said to mean that $p$ if $\alpha$ being true in some model is enough to guarantee that $p$ is the case with respect to that model. That is, $\alpha$ is said to mean that $p$ if in all the models in which $\alpha$ is true it is the case that $p$ (without requiring as well that $\alpha$ be true in only those models with respect to which it is the case that $p$). For instance, in this looser sense we could say that (g) means that there are infinitely many things

$$\forall x\forall y(fx=fy \to x=y) \land \exists x\forall y\neg fy=x$$

Any model in which (g) is true has a domain with infinitely many objects. Nevertheless there might be models with an infinite domain but where (g) is false.

Be it as it may, these two senses of a sentence of a formal language meaning something that we have considered in this section (2) are not the senses that interest us the most here. We have considered them just not to confuse them with the ones we do have a primarily interest in.

(3) Sometimes we take a formal language to be just like a regimented version of a part of natural language. We might, for instance, regard $a$ just as another name for David Armstrong, and to take $P$ to make the same contribution to the meaning of the sentences where it appears as philosophizes makes to the meaning of the English sentences where it appears, and to take $Pa$ just as another way (in addition to Armstrong philosophizes, Armstrong filosofa, and many
others) of expressing that Armstrong philosophizes; and we might take \( \exists x P x \) just as one alternative way of expressing that there is a thing that philosophizes. A formal language is one kind of language artificially created, just like Esperanto is.

Understood in this way then the sentence \( Pa \) can mean the same as the English sentence *Armstrong philosophizes*. The question now is, how do we manage to make a particular formal language, understood in the sense we are describing here, to mean what it means? When considering the sense in subsection (1) of a sentence of a formal language meaning something, we saw that providing a model and a truth theory in the standard way was not enough to have a formal language whose sentences would possess the characteristic that we are considering here: to mean the same as some sentences of a natural language. It might be thought, though, that we can obtain a language with such a characteristic if we amend the truth theory we were considering in (1) so as to avoid the features that were the basis for our argumentation that the sentences of \( \mathcal{L} \) did not mean the same as any English sentence.

We could avoid having the value of a predicate symbol to be a set by not assigning a value to it through the model but rather having one clause in the definition of satisfaction for each predicate symbol, this clause being of the same sort as the one we offer here for \( P \):

if \( \alpha = Pt \), where \( t \) is a term, then \( M \) and \( s \) satisfy \( \alpha \) iff \( M/s(t) \) philosophizes
Then we would have as a consequence:

\[(h) \quad P_a \text{ is true iff Armstrong philosophizes}\]

We could as well decide to use a stronger biconditional, instead of the one involving only material implication. There are several possibilities here, since conditionals can be postulated to be more or less strong\(^3\). For the sake of the argument let’s suppose we chose the strongest possibility and make the biconditional to be metaphysically necessary equivalence (we can think of it as placing a necessary operator in front of the whole biconditional sentence in (h)). This biconditional would intuitively be too strong since would require \(P_a\) to be true in any world where Armstrong philosophizes, even in those where \(P_a\) does not have its actual meaning, and means something which is not the case in that world. Nevertheless, this strong biconditional would still be too weak to avoid the difficulty (i) pointed out in subsection (1): we could still have another theory with respect to the same model such that all expressions would be assigned the same value and all sentences declared as true in exactly the same possibilities but that yields a biconditional where \(P_a \text{ is true}\) is not paired with \(Armstrong \ philosophizes\) but with another sentence that intuitively has another meaning. For instance, this alternative theory could yield (j)

\[(j) \quad P_a \text{ is true iff the square of 11 is 121 and Armstrong philosophizes}\]

---

\(^3\) We could require that the equivalence holds in all worlds where Edinburgh is the capital of Scotland, or in all physically possible worlds, in all worlds where the interpretation of language \(L\) is the same, etc.
Both the theory that would have (h) as a consequence and the theory that would have (j) as a consequence would assign the same value to all expressions and declare all sentences true or false in exactly the same circumstances. They are indistinguishable with respect to what they say about the language $\mathcal{L}$. So if we claim that according to one of the theories $Pa$ means that $p$ then we should also maintain that according to the other $Pa$ means that $p$. On the other hand, though, we would like to claim that according to the first theory, and given (h), $Pa$ means that Armstrong philosophizes, but then, given (j), we should claim that according to the second theory $Pa$ means that the square of 11 is 121 and Armstrong philosophizes. We are led, then, to the contradiction that the sentence $Pa$ does and does not mean the same according to the two theories. The contradiction seems to arise from supposing of each of the theories that it suffices to endow $Pa$ with certain specific meaning, in the same way that English sentences have meaning.

Even if the kind of theories considered so far do not suffice to make the sentences of $\mathcal{L}$ to mean in the sense that we are considering in this section (3), there is another way of accomplish it, which seems to be what we, one way or other, in fact do when introducing a formal language which is used and understood in the way we are considering here. This other way is to provide a translation from the formal language into a natural language. Unlike what was the situation in subsection (1) here there is no standard way to proceed, since the translation procedure is not usually presented in an explicit way. One possible way of interpreting the expressions of $\mathcal{L}$ by explicitly indicating how to translate them into English would be the following:

4 Regarding the sign "~" see footnote 9 in chapter 2.
(If $\alpha$ translates as $\beta$, we will also write $tr(\alpha)=\beta$)

"a" translates as "Armstrong"

if $v$ is a variable, $v$ translates as $v$

"P" translates as "philosophizes"

if $t_1$ and $t_2$ are terms, "\(t_1=t_2\)" translates as $tr(t_1)\sim tr(t_2)\sim$ is identical with "\(\sim tr(t_2)\sim"

if $\Theta$ is a predicate and $t$ is a term, "\(\Theta t\)" translates as $tr(t)\sim tr(\Theta)\sim$

if $\alpha$ is a formula, then "\(\sim \neg \alpha\)" translates as "it is not the case that "\(\sim tr(\alpha)\sim"

if $\alpha$ and $\beta$ are formulas, "\(\alpha \land \beta\)" translates as "it is both the case that "\(\sim tr(\alpha)\sim" and that ”\(\sim tr(\beta)\sim"

if $\alpha$ is a formula and $v$ is a variable, "\(\exists v \alpha\)" translates as "there is an object "\(\sim v\sim" such that "\(\sim tr(\alpha)\sim"

So we have, for instance, that the sentence of \(\mathcal{L}\) \(\exists x (\neg Px \land x=a)\) translates into English as *there is an object x such that it is both the case that it is not the case that x philosophizes and that x is identical with a.* And, of course, $P\alpha$ translates as *Armstrong philosophizes.* Since we understand the English sentences we understand what the meaning that we postulate for the sentences of \(\mathcal{L}\) is.

One comment about how formal languages are sometimes, and maybe even often, taught in introductory courses to logic which I believe has a significance beyond the pedagogy of logic: when formal languages and, in particular, first-order formal languages are first introduced, it is common to begin by explaining what sort of things can be expressed with these languages. So,
for instance, students are taught that *John loves Mary* can be expressed as $L_{jm}$, or that *there is something that loves everything* is expressed in a first order language as $\exists x \forall y L_{xy}$. That is: the semantics for the formal language is presented in the sense of (3): students are told what the expressions and the sentences of particular formal languages mean by giving English equivalents, and they are trained in translating from English into a formal language and from the formal language into English. Then they are told something of the following kind: 'Now we are going to do in a rigorous way what we have done so far in an intuitive way'. And then they are introduced to models, interpretation functions, assignments and the recursive definition of satisfaction. That is, they learn how to interpret a formal language in the sense of (1). Notice, though, that, whatever reasons there might be for presenting the topic in this way, the teacher who proceeds in this way is in some respect fooling her students: to interpret a language as in (1) is not a rigorous way of doing what we do when we interpret it as in (3). It is to do something else. We can see this in the fact that the sentences do not mean the same. To use our example once more: in the case of $\mathfrak{L}$, the sentence $Pa$ can mean that Armstrong philosophizes when interpret as in (3) but not when interpreted as in (1).

At this point we can introduce some terminology that will distinguish among different senses of what we have so far ambiguously called *a formal language*, or simply *a language*. We will refer to what sometimes is called *uninterpreted language* (which, if it does not have any semantic property would seem not to deserve the name *language* at all) as *a system of forms*; the only systems of forms we will be concerned with here are those of standard propositional logic, first-order and second-order logic, so that when we say 'a system of forms' we mean one the
those three types; we will refer to a system of forms with an interpretation in the sense of (1) as a formal language, and we will refer to a system of forms with an interpretation in the sense of (3) as a regimented language.

(4) Sometimes the sentences of a formal language or a regimented language that have some specific interpretation are said to mean in addition also something else because we are employing encodement: the objects the language talks about can have other objects associated with them, and some formulas of the language can be seen as codifying or playing the role of predicates about these other objects.

The most significant kind of encodement is the so called Gödelization where we codify the primitive symbols of the language (system of forms) of arithmetic by means of natural numbers. One way of doing it is, for example, to associate the numbers 1,3,5,7,9,11,13 and 15 to, respectively the constant 0, the monadic function symbol s, the 2-place function symbols + and ., and the primitive logical symbols $\exists$, $\neg$, $\land$ and $\equiv$. To the constant $x$, we assign the number $2i+17$. Furthermore we can assign an (even) number to each sequence of primitive signs of the language of arithmetic, and we also assign an (even) number to each finite sequence of finite sequences of primitive symbols of the language. Then when we interpret the language of arithmetic in the usual way as about natural numbers, we can see the sentences of the language

\footnote{We could do this using a 'pairing function' and Gödel's $\beta$-function. See, for instance, George Boolos' The Logic of Provability, (1993), Cambridge University Press, pp. 17 ff.}
of arithmetic as also encoding claims about the symbols and sequences of symbols of the
language of arithmetic, and to see formulas of arithmetic with n-free variables as encoding
predicates about the language of arithmetic. A very simple and uninteresting example would be
the formula \( \neg x = sss0 \) which because in its usual interpretation is true when the value of \( x \) is not
3 (and because the number that corresponds to the function symbol \( s \) is 3), can be seen as
encoding the predicate 'it is not the symbol \( s \).

When all the objects that are the interpretation of the language are encoding some object
or other (in the example of the previous paragraph: if every natural number encodes some
primitive symbol of the language of arithmetic, or some sequence or sequence of sequences of
primitive symbols) then instead of 'encodement' we could directly talk of just 're-interpretation'
of the language: the encoding provides us with another way of giving an interpretation for the
language. In the case of the language of arithmetic and the encoding of expressions and
sequences of expressions of the very language of arithmetic by means of natural numbers, we
could understand the codification as providing another interpretation of the language of
arithmetic: the domain of this alternative interpretation has as individuals the primitive symbols
of the language of arithmetic, sequences of those, and sequences of those sequences; and we
would give an interpretation for the different primitive non-logical symbols making use of the
correspondence we have between the elements of the domain of the usual interpretation (natural
numbers) and the elements of the domain of this alternative interpretation (symbols, sequences
of symbols, and sequences of sequences of symbols). For instance, according to the alternative
interpretation we would interpret the function symbol \( s \) as the function that assigns to an element
of the domain e (a symbol of the *language* of arithmetic, or a sequence of symbols, or a sequence of sequences of symbols), the symbol or sequence that is associated with a number which is the successor of the number associated with e (i.e., if e is associated with the number \( n \), \( n' \) is the successor of \( n \), and \( e' \) is the symbol or sequence associated with \( n' \), then the interpretation of \( s \) assigns \( e' \) to e).

We could distinguish different levels in how a formula or sentence that has certain interpretation about certain objects 'says' something or codifies some claim about some other objects. For instance we can say of a certain formula of the *language* of arithmetic with one free variable that codifies the predicate 'to be a formula provable in the theory Z' (or that it says that the value of the variable x is a formula provable in Z), only because on the usual interpretation of the *language* of arithmetic the formula is true for exactly those values of x that are numbers that are associated by the codification with sequences of primitive symbols of the *language* of arithmetic that are formulas that can be proved in Z. We could have stronger reasons, though, for saying of a formula \( \alpha(x) \) of the language of arithmetic that it *expresses* or *corresponds to* the predicate 'to be a formula provable in Z'. It could be that the formula \( \alpha(x) \) not only is true for exactly those values of x that are numbers that correspond to formulas that are provable in Z, but also that \( \alpha(x) \) is built up from subformulas that are coding the predicates 'to be a formula of the *language* of arithmetic', 'to be an axiom of Z', 'to be a sequence', 'to be a member of a sequence', 'to be earlier in a sequence than', 'to be the result of applying

\[ 6 \] Z is the theory of Elementary Peano arithmetic. For a list of its axioms see, for instance, Boolos, op. cit., pp. 18-19.
Modus Ponens to', and \(\alpha(x)\) can be seen as saying that 'there is something that is a derivation of \(x\) from the axioms of \(Z\)' or more explicitly 'there is something that is a sequence of formulas such that each one is either an axiom of \(Z\) or the result of applying Modus Ponens to two earlier formulas in the sequence, and the last formula of the sequence is \(x\)'. The subformulas of \(\alpha(x)\) can, in turn, be built up from other subformulas that are also coding predicates about the language of arithmetic (for instance the formula \(\beta(x)\) corresponding to 'to be a formula' can contain subformulas that correspond to the predicates 'to be an atomic formula', 'to be the negation of', 'to be the conjunction of', 'to be an existential quantification of', and \(\beta(x)\) can be seen as saying that 'there is a sequence such that each member is either an atomic formula, or the negation of an earlier member, or the conjunction of two earlier members, or an existential quantification of an earlier member, and \(x\) is the last member of the sequence'). It seems clear that one such formula \(\alpha(x)\) can be said in a more proper or fuller sense that expresses the predicate 'to be provable in \(Z\)' than one formula that simply is true for the right values of \(x\).

There would be a lot more to say about encodement and Gödelization (in particular, it would be interesting to clarify what exactly the distinction between 'expressing in a more/less full sense' consists in). We leave it here, though, since the sense of a sentence having certain meaning that we have considered in this section is not one we wish to focus in in this article, and we have included it just for the sake of completeness and to distinguish it from the other senses.
2. Natural language and regimented languages

What is the relationship and the differences between natural language and regimented languages? In this section we are going to briefly look at three or four aspects of this relationship. Some of our comments will be tentative or inconclusive. This is a difficult topic. One point of considering here the relationship between natural language and regimented languages is simply to see what sort of issues arise regarding this relationship and to be able to separate them from the issues that arise about the relationship between formal languages and regimented languages.

As we saw a regimented language is interpreted by using part of natural language. A regimented language, though, has some differences even with that part of natural language that is used to interpret it. One of them has to do with the domain of quantification. When we introduced one particular regimented language in section 1-(3) we did not provide any specific domain of quantification --we took our quantifiers to range over everything there is. We already mentioned that, unlike what is the case for formal languages, there is no standard way of characterizing regimented languages. When they are provided, though, a domain of quantification is usually specified, i.e., there are some things which are stipulated to be the ones the language will talk about. So if the domain of quantification consists of the things that are p, then an existential quantification $\exists x$ will be translated as ‘There is a thing x which is p such that’. The domain of quantification is the same for all the sentences of a particular regimented language. This is not the case with respect to natural language, where the domain of quantification can change from one sentence to another, or even from one part of a sentence to another, as in (k) and (l):

165
(k) I entered the room. Everything was in order. I looked into the fridge. Everything had been stolen.

(l) After the attack on Ganymede, someone was happy to learn that everyone was dead.

In (k) what has been stolen is not what was claimed to be in order, and in (l) whoever was happy after the attack is not someone who was dead.

Having a fixed domain for all uses of quantifiers is, then, a feature that distinguishes a regimented language form natural language. This feature of regimented languages is one aspect of two related general characteristics of these languages: not to be subject to context dependency and to approximate the ideal of displaying in an explicit way all the features that are relevant for the meaning of the sentence.

Another such feature is that in a regimented language different syntactic categories correspond to different semantic categories. We will not go here into the very interesting and very difficult topic of characterizing what a semantic category is. We will make just one comment regarding predicate symbols. Predicate symbols are translated by means of expressions whose meaning is such that either applies or does not apply to an individual (in the case of a monadic predicate) or to n individuals (for an n-place predicate). We might, for instance, interpret the 1-place predicate symbol $R$ by indicating that it translates as *runs*. Then we would have that
it applies to those individuals that run and does not apply to those individuals that do not run. Now: we could also interpret the monadic predicate symbol $Q$ by indicating that it translates as runs quickly. If we translate the constant $a$ as Armstrong, do $Ra$ and $Qa$ mean, respectively, the same as Armstrong runs and Armstrong runs quickly?

Our concepts of 'meaning' and 'meaning the same' are probably not precise enough as to allow us to go into too fine-grained distinctions, but I would just want to point out that it is not obvious that the answer to the question is 'yes'. Armstrong runs follows logically from Armstrong runs quickly, whereas it would seem that $Ra$ does not follow logically from $Qa$ (we are here appealing to the intuitive, pre-theoretic notion of 'following logically from' or 'being a logical consequence of'). If we believe that logical properties depend only on the meaning of the expressions (and not, for instance, on their spelling or pronunciation), then the two pairs of sentences must differ in meaning.

We might wonder: given the interpretation that $Ra$ and $Qa$ have, is it really the case that $Ra$ is not a logical consequence of $Qa$? I think that it's clear that $Ra$ is an analytic consequence of $Qa$ (or that $\neg(Qa \land \neg Ra)$ is an analytic truth). Nevertheless, it seems reasonable to believe that $Ra$ is not a logical consequence of $Qa$, in the same way that we believe that the English sentence John is an unmarried man is not a logical consequence of John is a bachelor. This situation regarding sentences of a regimented language and sentences of English does not arise because of any special features of regimented languages. In Catalan a donkey is called a ruc and a young donkey is called a pollf; the adjective jove means the same as young. The Catalan sentence
Francis és un pollí would seem to mean exactly the same as the English sentence Francis is a young donkey (if Catalan speakers did also speak English they could have introduced the word pollí into their language by postulating while using English 'from now on in Catalan with pollí we are going to mean a young donkey'). But while the English sentence Francis is a donkey is a logical consequence of Francis is a young donkey, the Catalan sentence Francis és un ruc is not a logical consequence of Francis és un pollí. We might want to say that the meaning of a sentence is structured and that two sentences can 'have the same meaning' because they are 'describing the same situation' but, nevertheless, mean something with different structure. This is what seems to be the case with respect to the sentences of a regimented language Ra and Qa considered above, and the English sentences that we used to interpret them.

Notice that the situation would have been different if the sentences of a regimented language were regarded as notational simplifications or abbreviations of English sentences. We could have introduced such type of language by means of some clauses that would look very much like the ones we gave to introduce our sample regimented language in 1-(3): all we would need to do would be to substitute "is an abbreviation of" for "translates as". If "Ra" abbreviates "Armstrong runs", "Ra" means something only through its connection with "Armstrong runs".

---

7 If we want steer clear of the difficulties posed by adjectives like fake, we could consider instead the fact that Francis is a young donkey is a logical consequence of Francis is a young donkey, but Francis és un ruc jove is not a logical consequence of Francis és un pollí.

8 It would, of course, be an important and interesting task to try to make more precise what we mean by saying that two sentences describe the same situation. We will not attempt to do it here.

9 See also footnote 17 and, in general, subsections 4.2(f)-(g) in chapter 2.
This relation of 'being an abbreviation of' can not simply be analyzed in terms of those of 'being a name of', 'being a token of' and 'being a type of'. If "Ra" is an abbreviation of "Armstrong runs" then it is not the case that "Ra" names "Armstrong runs", since to use "Ra" is not to mention "Armstrong runs" but rather to express something about Armstrong; it is not the case either, though, that when we make a particular use of "Ra" we have used a token of the sentence "Armstrong runs": we have only used a token of the expression "Ra". If Ra and Qa are just abbreviations of Armstrong runs and Armstrong runs quickly then certainly the same logical relations must hold between Ra and Qa, and the fully expanded English sentences.

Another difference between regimented languages and a natural language like English arises because the use of regimented languages is not subject to the conversational norms that in the case of natural languages have an influence on what is communicated with the use of some sentence. For instance, if I say There is a man waiting for John it will be understood that there is only one man who is waiting, whereas the sentence $\exists x W x j$ (with the appropriate interpretation of W and a, and quantifying over men) can not be taken to communicate that there is only one man waiting for John. This is not because the existential quantifier in the regimented language has a different interpretation from the existential quantifier in natural language (this could hardly be the case given that we have interpreted existential quantification in the regimented language by means of existential quantification in English and problems such as the ones regarding the structure of meaning that we mentioned above do not arise here). The use of the English sentence There is a man waiting for John is subject to the effect of the Cooperative Principle and, in
particular, to the maxim of quantity: the maxim requires to give as much information as might be needed; whether there is only one man or more would usually be relevant information in a context where the sentence is used; lacking information to the contrary it will be assumed that the person using the sentence knows whether there is only one man or there are more; that there are two men, that there are three men, that there are several men or that there are many men can be expressed with as much brevity and simplicity as that there is a man; the speaker is abiding by the Cooperative Principle and chose to say that there is a man rather than any of the other stronger claims, this must be because he knows the stronger claims not to be true and, so, there must be only one man. A reasoning of this kind is what makes it the case that when we use *There is a man waiting for John* we usually communicate that there is only one man waiting for John.

Similar observations could be made with respect to other contrasts between sentences of a regimented language and natural language sentences that are used to interpret them. (For instance: the effect of the maxim of manner and the difference between *John got rich and he took up philosophy* and *John took up philosophy and he got rich*, but the equivalence between \((Rj \land Tj)\) and \((Tj \land RJ)\) [with their obvious interpretation]).

This distinction between natural language and regimented languages seems to be not so much about the languages themselves but about their use. The sentences of a regimented language do not communicate anything other than what constitutes their meaning because they are not evaluated within the framework that is assumed when we consider the use of a sentence.
as part of the cooperative effort of a conversation.

So far we have been commenting on the relationship between regimented languages and that part of natural language that is used to interpret them. The sentences of a regimented language, though, can also 'express' or 'formalize' what it is expressed by sentences not belonging to that part of natural language that is used to interpret the regimented language. For instance we can formalize in a first-order regimented language where we quantify over humans and where $M$ translates as *is a man*, $P$ translates as *philosophizes* and $c$ translates as *Bill Clinton* the English sentence *All men philosophize* as $\neg \exists x(Mx \land Px)$ (or if we were to introduce two clauses that would interpret "∀" and "→" in the obvious way, we could 'formalize' the English sentence also by $\forall x(Mx \rightarrow Px)$). We could 'formalize' a sentence with negation inside the verb phrase like *Bill Clinton does not philosophize* as $\neg Pc$. A large part of English, though not all, could be 'formalized' by means of a first-order regimented language. Now, what does it mean that a sentence of a regimented language 'formalizes', 'corresponds to', 'symbolizes', 'expresses the same as' or 'represents' some sentence of English? I think that in different occasions people mean different things by 'formalize' (or any of the other expressions).

What it does not mean, it seems clear, is 'sameness of meaning' in an strong sense: a sense such that, for instance, according to it to mean that some men do not philosophize is not the same as to mean that it is not the case that there is a thing $x$ such that it is both the case that $x$ is a man and that it is not the case that $x$ philosophizes. If we assume such a strong sense of meaning the same, then to provide a formalization can not consist in providing a sentence with
the same meaning or otherwise it would not even be possible to formalize *Some men do not philosophize* as $\exists x(Mx \wedge \neg Px)$.

Sometimes to 'formalize' an English sentence in a regimented language is understood so that it amounts to giving the *logical form* of the English sentence. I will not try here to explain what giving the logical form of an English sentence consists in (this is a difficult question), let just point out that it involves indicating what kind of expressions there are in the English sentence (i.e. the semantic category of the different expressions in the sentence) and how they are combined; to formalize in a regimented language is a particularly suitable way of indicating the logical form of an English sentence, since the sentence in a regimented language that formalizes the English sentence shows in a clear and explicit way what the different semantic categories of the expressions are since they correspond to the different syntactic categories.

Sometimes we speak of formalizations in what seems to be a more relaxed sense: different sentences of a regimented language can all be regarded as good formalizations of an English sentence provided that the regimented sentences are logically equivalent (even if we might also want to say that some of them are more 'natural' formalizations than the others). So for instance, given a regimented language where $K$ translates as *is king of France*, and $B$ translates as *is bold*, we can consider each of (n), (o) and (p) as good formalizations of (m)

(m) The king of France is bold

(n) $\exists xKx \land \forall x\forall y(Kx \land Ky \rightarrow x=y) \land \exists x(Kx \land Bx)$
This sense of 'formalization' seems to be the one that is appealed to when claiming that the English sentence *Some critics admire one another and no one else* is not only formalizable by means of the second order formula (q), but also by means of the first order formula (r)\(^{10}\), in a regimented language where A translates as *admires*:

\[\begin{align*}
(q) & \quad \exists x \forall y ((Xx \land Xy \land x \neq y) \land \forall x (Xx \rightarrow \forall y (Axy \leftrightarrow \{Xy \land y \neq x\}))) \\
(r) & \quad \exists y Azy \land \forall x ((z = x \lor Azx) \rightarrow \forall y (Axy \leftrightarrow \{(z = y \lor Azy) \land y \neq x\})))
\end{align*}\]

It might seem that in this sense of 'formalization', what is required of the formalization of an English sentence by means of a sentence of a regimented language is that (at least some of) the logical properties (logical truth, and the relation of logical consequence) are (in some sense to be made precise) preserved.

Two questions: First, how can we make clear and explicit this notion of 'preservation of the logical properties when an English sentence is formalized by means of a sentence of a regimented language'? It is not obvious how to make this notion precise. Notice, for instance, that this requirement of preservation of logical properties does not involve just one single English

\[\begin{align*}
(o) & \quad \exists x \forall y ((K_y \leftrightarrow x = y) \land Bx) \\
(p) & \quad \exists x \forall y ((K_y \rightarrow \neg (y = x \rightarrow \neg B_y)) \land Kx)
\end{align*}\]

\(^{10}\) See Boolos' *To be is to be the value of a variable (or some values of some variables)*, The Journal of Philosophy 81, (1984) p. 439.
sentence and one single regimented sentence, but rather 'many' English sentences and 'many' regimented sentences. Second, does one such requirement of preservation of logical properties really provide a necessary and sufficient condition for a sentence of a regimented language to be a correct formalization in the sense in which we believe that, say, (n) or (o) are correct formalizations of (m)? Answering the first question will, of course, provide the basis for trying to address the second one.

I think we can make precise the idea of 'preservation of the logical properties when formalizing' in the following way. We will say that a function F from a set of English sentences A into the set B of all the sentences of a first-order regimented language $\mathcal{L}$ is interesting if conditions (C1) and (C2) hold

(C1) F upholds translation (by this we mean that if $s$ is an English sentence that, according to the translation procedure that gives the interpretation of $\mathcal{L}$, is the translation of some sentence $s'$ of $\mathcal{L}$, then $s \in \text{dom}(F)$ and $F(s)=s$).

(C2) If $\Sigma \subseteq B$, $\alpha' \in B$, $\Sigma' = \alpha'$, $F(\alpha) = \alpha'$ and $F[\Sigma] = \Sigma'$, then $\Sigma = \alpha$

Here we use the expression $\Sigma = \alpha$ simply as an abbreviation of $\alpha$ is a logical consequence of $\Sigma$.

We will presently make some comments on our definition of F. Let's already state, though, how this definition of a function being interesting attempts to make precise the notion
of 'formalization that preserves the logical properties': If a function is interesting then it is a
function that preserves the logical properties (as we will see in comment (v) below, an interesting
function might not preserve all the logical properties); regarding a single sentence we can now
say the following: a sentence $s'$ of a regimented language preserves the logical properties of an
English sentence $s$ iff there is an interesting function $F$ such that $F(s) = s'$. Notice that in this
precise sense of 'a sentence of a regimented language preserving the logical properties of an
English sentence' we have just introduced if a function is interesting then all the sentences in its
range preserve the logical properties of the English sentences of which they are the value.

Five observations on our definition of $F$:

(i) Notice that $F[\Sigma] = \Sigma'$ iff $\Sigma \subseteq F^{-1}[\Sigma']$ and for each $\sigma' \in \Sigma'$ there is a $\sigma \in \Sigma$ such that $\sigma \in F^{-1}[\{\sigma'\}]$.

(ii) If $F$ is interesting and $F(s_1) = s = F(s_2)$ then $s_1$ and $s_2$ must be logically equivalent. Justification:

$\{s\} = s$, $F(s_2) = s$ and $F[\{s_1\}] = \{s\}$, so by (C2) $\{s_1\} = s_2$; analogously $\{s_2\} = s_1$.

(iii) Condition (C1) ensures that there are no 'permutations of non-logical symbols'. By this we
mean the following: Suppose that $\mathcal{L}$ has two predicates $K$ and $B$ interpreted as is king of France
and is bold; if we only had condition (C2), then if $F$ is interesting so would be a function $F'$
which is such that if $F(s) = \alpha$ then $F'(s) = \alpha$', where $\alpha'$ comes from $\alpha$ by substituting all occurrences
of $K$ for $B$, and all occurrences of $B$ for $K$; since the sentences of $\mathcal{L}$ are interpreted, though, it can
not be that both $F$ and $F'$ are providing intuitively good formalizations of the English sentences
We could have dispensed with (C1) if we had taken a somewhat different approach. What we have done is to assume that we start off with an (interpreted) regimented language and then we look for conditions that ensure that each English sentence in the domain of the function gets assigned a correct formalization (i.e. a sentence of the regimented language that means 'the same' as the English sentence in some weak sense of 'meaning the same' --two logically equivalent sentences mean the same in this weak sense). An alternative way to proceed would have been to start off with an (uninterpreted) first-order system of forms and look for a condition on the function that ensures the following: that it is possible to interpret the system of forms (i.e. to make it become a regimented language) in such a way that each English sentence in the domain of the function gets assigned a correct formalization. If we had proceeded in this second way there would be no need to prevent 'permutations of non-logical symbols', and so we might think that we do not need condition (C1) (see also, though, observation (iv)). Also, if we proceeded in this way, in (C2) we would not be using the intuitive pre-theoretic notion of logical consequence, but rather the model theoretic notion (of which we will talk in the next section). This would not be problematic since it is undisputed that this model theoretic notion is at least extensionally correct.

(iv) Condition (C1) also ensures that the set of English sentences in the domain of function F is 'big enough'. If the set A had, say, just two English sentences then F could be interesting and that would not yet give us any guaranty that F assigns to each of the English sentences in A a
sentence of the regimented language that correctly formalizes it. Condition (C1) ensures that F will be a function onto B. If we did not have (C1) we could introduce this effect of condition (C1) by directly requiring F to be a function from A onto B.

(v) Should we also have a condition such as (C3)?

\[(C3) \quad \text{If } \Sigma \subseteq A, \alpha \in A, \Sigma \models \alpha, \text{ then } F[\Sigma] = F(\alpha)\]

We might think that this condition would place too strong a restriction on F. It would not allow us, for instance, to have both *John runs* and *John runs quickly* among the English sentences in the domain of the function. Still, if we wanted to restrict the English sentences that can be in the domain of the function to those all of whose logical properties can be captured in a first-order regimented language, then we should certainly have (C3) in addition to (C1) and (C2).

Let's now consider the second question we posed above. If \( s' \) is a sentence of a regimented language that preserves (in the sense we have introduced) the logical properties of the English sentence \( s \), is \( s' \) a correct formalization of \( s \), in the same sense in which, say, (n) and (o) are good formalizations of (m)? I don't believe it is, though I do not have conclusive evidence that it is not. All what I will do here is to try to show that the requirement of preservation of logical properties is much weaker than one might at first have thought.

That a sentence of a regimented language \( s' \) preserves the logical properties (in our sense)
of an English sentence $s$ is certainly a necessary condition for $s'$ to be a correct formalization of $s$.\(^{11}\) I want to argue that it is not clear that it is also a sufficient condition. I believe this is so even if we add condition (C3) to our definition of an interesting function. Actually, for the purposes of our discussion of the sufficiency of preserving the logical properties for being a good formalization we can suppose that our definition of an interesting function includes (C3).

We will show the following: if $\mathcal{L}$ is a regimented language, and $F$ is an interesting function whose domain is a set of English sentences $A$, and whose range is the set of sentences of $\mathcal{L}$ then there is a function $G$ whose domain is $A$, that satisfies (C2) and (C3) and whose range is a set of sentences of the language of propositional logic. (Notice that we are not claiming that $G$ also meets (C1)). This result indicates that meeting conditions (C2) and (C3) is not sufficient for ‘capturing’ any property of the English sentences that depends essentially on the predicational structure or on the quantificational structure. This fact would seem, in turn, to take off plausibility from the idea that every interesting function assigns correct formalizations to the English sentences in its domain.

Let’s prove the claim stated in the previous paragraph. We will use the following result: Let $Q$ be the set of sentences of some language for quantificational logic. (A language for

\(^{11}\) Notice, though, that (C2) prevents formalizing (i) as (ii) (and, in particular, for instance, prevents formalizing the antecedent of (i) as the antecedent of (ii)), since (ii) is a logical truth whereas (i) is not:

(i) \[ \text{If there are at least three things then there are at least two things} \]
(ii) \[ \exists x \exists y \exists z (x \neq y \land x \neq z \land y \neq z) \rightarrow \exists x \exists y (x \neq y) \]
quantificational logic is a first order system of forms that has a denumerable number of non-
logical primitive symbols of each category [constants, n-place function symbols (for each n), n-
place predicate symbols (for each n)], so that any specific first order system of forms can be seen
as a subset of this language for quantificational logic). Let P be the set of sentences of some
language for propositional logic. (A language for propositional logic contains a denumerable
number of sentence letters); then there is a on-one function H with domain Q and range P such
that: for any $S \subseteq Q$ and $\alpha \in Q$, $S = \alpha$ iff $H[S] = H(\alpha)$. This result is due to George Boolos\(^{12}\).

Let $F$ be an interesting function from a set $A$ of English sentences into the set $B$ of all
the sentences of a regimented language $\mathcal{L}$. Let $Q$ be the set of sentences of some language for
quantificational logic that includes the sentences of $\mathcal{L}$. Let $P$ be the set of sentences of a language
for propositional logic. Let $H$ be a 1-1 function from $Q$ onto $P$ such that: for any $S \subseteq Q$ and $\alpha \in Q,$
$S = \alpha$ iff $H[S] = H(\alpha)$ (We know that $H$ exists by Boolos’ result). Let $H^\downarrow$ be the restriction of $H$ to
B. We have that: $H^\downarrow$ is a one-one function from $B$ onto $H[B]$ and that

\[(*) \text{ for any } S \subseteq B \text{ and } \alpha \in B, \text{ } S = \alpha \text{ iff } H^\downarrow[S] = H^\downarrow(\alpha).\]

\(^{12}\) George Boolos (1970?), 'Quantificational and Propositional Logic are Isomorphic'.
Unpublished manuscript, MIT, Cambridge. (Regarding the question mark on the date: no one I
have asked, including the author himself, has been able to recall when exactly this short paper
was written). The proof of the result has the following structure: We show that the Lindenbaum
algebras of $Q$, and of $P$ are atomless. We apply the well known theorem that says that any two
atomless Boolean algebras are isomorphic. Given an isomorphism between the two algebras we
define an isomorphism between $Q$ and $P$.

I am indebted to Richard Cartwright for letting me know about the existence of Boolos’
paper and for pointing out to me its relevance for this discussion.
Let $G$ be the composition of $F$ and $H^\downarrow$. $G$ is a function from $A$ into $H^\downarrow[B]$, i.e., into a set of sentences of propositional logic. Let's see that $G$ satisfies (C2) and (C3)

(C2): Suppose that (1) $\Sigma \subseteq H^\downarrow[B]$, (2) $\alpha' \in H^\downarrow[B]$, (3) $\Sigma' = \alpha'$, (4) $G[\Sigma] = \Sigma'$ and (5) $G(\alpha) = \alpha'$.

We want to see that $\Sigma = \alpha$.

Given (1), $H^\downarrow^{-1}[\Sigma'] \subseteq H^\downarrow^{-1}[H^\downarrow[B]]$, and since $H^\downarrow$ is one-one $H^\downarrow^{-1}[H^\downarrow[B]] = B$, so we have (1)* $H^\downarrow^{-1}[\Sigma'] \subseteq B$. Given (2), $H^\downarrow^{-1}(\alpha') \in H^\downarrow^{-1}[H^\downarrow[B]]$, and so we have: (2)* $H^\downarrow^{-1}(\alpha') \in B$. Given (1)*, (2)* and the fact that $H^\downarrow$ satisfies (*) we have that: $H^\downarrow^{-1}[\Sigma'] = H^\downarrow^{-1}(\alpha')$ iff $H^\downarrow[H^\downarrow^{-1}[\Sigma']] = H^\downarrow[H^\downarrow^{-1}(\alpha')]$.

Since $H^\downarrow(H^\downarrow^{-1}[\Sigma']) = \Sigma'$ and $H^\downarrow(H^\downarrow^{-1}(\alpha')) = \alpha'$ we have: $H^\downarrow^{-1}[\Sigma'] = H^\downarrow^{-1}(\alpha')$ iff $\Sigma' = \alpha'$. Given (3) we have (3)* $H^\downarrow^{-1}[\Sigma'] = H^\downarrow^{-1}(\alpha')$.

Given (4), $H^\downarrow(F[\Sigma]) = \Sigma'$, since $H^\downarrow$ is one-one we have (4)* $F[\Sigma] = H^\downarrow^{-1}[\Sigma']$. Given (5), $H^\downarrow(F(\alpha)) = \alpha'$. Since $H^\downarrow$ is 1-1, we have (5)* $F(\alpha) = H^\downarrow^{-1}(\alpha')$.

Given (1)*, (2)*, (3)*, (4)*, (5)* and the fact that $F$ satisfies (C2) we have that $\Sigma = \alpha$.

(C3): Suppose that $\Sigma \subseteq A$, $\alpha \in A$ and $\Sigma = \alpha$. We want to see that $G[A] = G(\alpha)$. Given that $\Sigma = \alpha$ and since $F$ satisfies (C3) we have $F[\Sigma] = F(\alpha)$. Given that $H^\downarrow$ satisfies (*), $H^\downarrow[F[\Sigma]] = H^\downarrow(F(\alpha))$, and so $G[\Sigma] = G(\alpha)$. Q.E.D.
We have shown, then, that if there is an interesting function, i.e., a function that preserves the logical properties of the English sentences in its domain, that has as a range all the sentences of a first-order regimented language, then there is another function that also meets (C2) and (C3) but whose range is a set of sentences of *propositional* logic. This suggests that preserving the logical properties is a weaker requirement than it might seem.

Notice, though, the following: the range of function $G$ might not be 'closed', that is, there might be sentence letters that appear in some formulas of $G[A]$ (i.e. $H[B]$) but that are not themselves members of $G[A]$; and there might be a formula $\alpha$ that appears in some formula of $G[A]$ but which is such that, for instance, $\neg \alpha$ is not in $G[A]$. We could have made a stronger case in showing the weakness of the requirement of preservation of logical properties if our function $G$ did not only meet (C1) and (C2) but also two other conditions that an interesting function meets: (a) it satisfies (a version of) (C1), and (b) the range of $G$ contains all the sentences of a particular propositional language (not just a subset). If $G$ met (b) then we would also have (a): we could interpret each sentence letter $s$ by giving as an English translation one (any of them if there is more than one) of the English sentences that $G$ assigns to $e$ (this requirement for sentence letters would already guarantee that more complex sentences meet the equivalent of (C1) for propositional calculus). Whether we can define a function that in addition to (C1) and (C2) also meets (b) is, at present, still an open question.
3. Regimented languages and formal languages

In this last section we will make some comments on the relationship between formal languages and regimented languages. In this way we will also indirectly talk about the more important relationship between formal languages and natural language. Having so to speak broken the relationship between formal languages and natural languages in two parts (natural language-regimented language and regimented language-formal language) will help, I hope, to make clearer that this relationship gives rise to issues that are of different kind and that should be distinguished.

What is the relationship between formal languages and regimented languages? I want to claim that formal languages are models of regimented languages. Here by model we do not mean what is meant in model theoretic semantics, i.e., an structure or interpretation, but rather what we usually mean when we say that we construct a model of something: something else that has some of the same properties as the original object, and that it is usually made in order to facilitate studying those properties that the two objects have in common. To avoid terminological confusions we will call a model in this (when not talking about logic) most common sense a modelation.

One fact that makes it easy to overlook that formal languages are not regimented languages, but just modelations of them, is that it is often very easy to go from a formal language to a 'corresponding' regimented language and vice versa. For instance given a formal language
where, say, the interpretation of the predicate symbol \( P \) in the model \( M=\langle D, F \rangle \) is given with a clause of the form in (s)

(s) \[ F(P) = \{ x: x \text{ so-and-so} \} \]

we have a corresponding way of interpreting the predicate symbol \( P \) in a regimented language, namely with a clause of the form in (t)

(t) \[ P \text{ translates as 'so-and-so'} \]

And conversely, given a clause of the form in (t) which allows us to interpret the predicate \( P \) in a regimented language, we can think of the corresponding way of interpreting \( P \) in a formal language by means of a clause like (s). The same could be said for the other kinds of expressions.

Even if in many cases an interpretation of a system of forms as a regimented language already suggests an specific way of interpreting the system of forms as a formal language, and also vice versa, the two sorts of 'language' are very different. Remember that as we saw at some length in section 1 a sentence of a formal language does not really mean anything in the way that sentences of natural language or of a regimented language do.
George Boolos writes \(^{13}\)

When we say that + denotes plus in \(N\), using "plus" or a synonym to say so, we allow it to be understood that + is to have the sense of "plus", whatever that might be (and not, say, that of "plus the cube root of the square root of the cube of the square of"). Similarly for the other symbols of the language, including the variables, the manner of specification of whose range, i.e., as over the natural numbers, contributes in large measure to the determination of the meanings of quantified sentences of PA.

In this passage professor Boolos seems to be aware of the difference between having a formal language (whose sentences would not really have any meaning) and having a regimented language (whose sentences can have the sort of meanings that we intuitively attribute, for instance, to the sentences of the 'language’ of arithmetic), and of the tension that arises between interpreting the 'language’ (system of forms) of arithmetic as a formal language while pretending at the same time that the sentences have meaning and say things about the natural numbers in the way that English sentences say things about the natural numbers. He seems to pretend to be having both a formal language and a regimented language when he introduces the standard formal language for the 'language’ of arithmetic. He suggests that when introducing a formal language we are also introducing a corresponding regimented language. Other authors also seem to assume something like this, even if they do not make it explicit in the way Boolos does. Doing so without saying anything else, though, is unjustified. If nothing else is added the definition of

\(^{13}\) op. cit. p. 33.
a formal language does not provide by itself anything else other than the formal language itself, and this sort of 'language', as we have argued, is not a regimented language.

We claimed above that a formal language is a modelation of a regimented language. In what sense is a formal language a modelation of a regimented language? At least in the following sense: a formal language models the way the truth of the sentences of a regimented language is affected by the combined effect of the meaning of the expressions of the regimented language and the way the world is. Here by 'the way the world is' we do not just mean 'the way the world actually is', rather we mean that the formal language models how one way the world might be and the meaning that the expressions of some regimented language have would affect the truth of the sentences of the regimented language. With respect to each specific expression of a regimented language, a formal language models how the meaning of the expression in the regimented language and a way the world is affect the contribution that the expression makes to the truth value of the sentences where the expression occurs.

I believe we can see more clearly what I take to be the way a formal language models a regimented language if we look at it from the perspective of the distinction that John Etchemendy draws between representational semantics and interpretational semantics. The truth of a sentence depends on two parameters: 'broadly speaking, the language and the world' (p. 18). If we keep the language parameter fixed and we give values to the expressions of a

---

language trying to reflect the effect that the different ways the world might be would have on
the truth of the sentences of the language we are engaging in *representational semantics*. An
interpretation or a model (here in the model-theoretic sense) would correspond to one way the
world might be. If we keep the world parameter fixed and give values to the expressions of the
language (respecting the semantic category of each expression) to reflect the different ways that
the different possible meanings of the expressions would affect the truth of the sentences of the
language we are engaging in *interpretational semantics*. When doing interpretational semantics
an interpretation or model corresponds to one meaning that the language can have\(^{15}\).

Etchemendy notes that sometimes the same interpretation or model can be seen from the
representational or from the interpretational perspective, but he insists that the two sorts of
semantics are two different sorts of enterprise and that often they do not overlap. He explains that
the usual way in which we give a model-theoretic semantics for a first order language (in our
terminology: the usual way we proceed when we provide a first order formal language) accords
only with viewing models from the interpretational perspective and not with viewing them from
the representational perspective. For instance if we have two predicates \(P\) and \(Q\) and we feel free
to consider a model where, say, both predicates get assigned the set \(\{1,2\}\), then we can not be
giving a representational semantics of a language where \(P\) means 'to be completely green', and
\(Q\) means 'to be completely yellow'.

\(^{15}\) Again, when respecting the semantic category of each expression. From now on when
talking about 'the different meanings' or 'interpretations' that 'an expression can have' it can be
assumed that we mean 'different meanings in the same semantic category'.
There is a third possibility regarding what we do when interpreting a first order system of forms by giving a model that Etchemendy does not consider. I contend that the right way to view what we do when we use model-theoretic semantics is to view it in accordance to this third possibility: each model corresponds *not* to one possibility of how the world might be (representational semantics) or to one possibility of what the language might mean (interpretational semantics), but rather to one possibility of the *combination* of the two factors, i.e. to one way the world might *be* and one meaning that the expressions of the language might have. Here we should add a qualification: in the standard way to proceed, these different possible interpretations of the expressions of the language that we consider with each model do not only have the constraint of respecting the semantic category of each expression but they also have the constraint of keeping fixed the value that assign to the so called *logical symbols*.

When we interpret, for instance, the system of forms that has just one predicate symbol $P$ with a model that has as a domain the set $\{1,2,3\}$ and that assigns $\{1,2\}$ to $P$ we are considering a way the world could be *be* and a meaning that the system of forms could have (a meaning in the way regimented languages have meaning) such that: the world has some individuals such that our (regimented) language talks about (has as a domain) three of them and, furthermore, the world is such that the meaning that $P$ has applies to two of the three individuals.

It might turn out (in fact this is what happens, for instance, for first order languages) that if, on the one hand, we consider just all the models that correspond to different meanings that the expressions of a system of forms could have *given that the world is as it is* (interpretational
semantics) and, on the other hand, we consider all the models that correspond to all the possible combinations of a way the world could be and a meaning that the expressions of the language can have we end up considering the same models. Still, even if (unlike what was the case with respect to representational vs. interpretational semantics) here there is no extensional difference between taking the interpretational approach or our 'third way' approach, there is an important conceptual difference between viewing what we do when doing model theoretic semantics as doing interpretational semantics or as doing semantics according to our 'third way' approach. This difference becomes crucial when we consider the use we can make of formal languages to study the fundamental logical properties --logical consequence and logical truth.

Our intuitions about these two fundamental properties (which because they are fundamental we are particularly interested in clarifying) seem to be roughly the following: a sentence (of a natural or a regimented language) is a logical truth if it is true just in virtue of the meaning of certain expressions, the so called logical expressions, and the 'form' of the sentence. 'Form' here does not mean 'grammatical form' but rather 'the semantic category of the different expressions of the sentence and the way the expressions are combined'. As for the logical expressions they are expressions characterized by having a meaning of a particularly general kind; unlike the other expressions, it is meaningful to apply logical expressions to all sorts of discourse. Besides this vague and general idea, our intuitions about logical expressions seem to include enough as to allow us to recognize one when we see it, with the limitations imposed by the vagueness that the concept of 'logical expression', as most others, surely has: 'or' and 'every' are logical expressions, but 'John' and 'dog' are not, we might not be completely sure regarding
expressions such as 'is one of them' or 'exactly five'. A sentence is a logical consequence of some sentences if: the sentence is true if all these other sentences are and this is so just in virtue of the meaning of the so called logical expressions and the 'form' of all the sentences involved. For simplicity we will from now on focus just on the notion of logical truth.

We are certainly interested in making these intuitions more precise. One way of doing so is by realizing that we capture these intuitions if we say that a sentence is a logical truth if it is true whatever way the world might be and whatever the meanings of the non-logical expressions might be, provided that they have a meaning that keeps them in the same semantic category. This formulation does not appeal to the notion 'in virtue of' which was certainly in need of clarification.

Having this formulation of what it is for a sentence to be a logical truth we might want to restrict our attention to regimented languages since they seem to be rich enough so that what we say about the logical properties as applied to them is extendable to a language in general but at the same time, since they are regimented, they avoid some of the unwelcome complications of natural language (some having to do with facts that we already mentioned in section 2, like the effect of Gricean maxims, others having to do with some facts about natural language such a.. the vagueness about which expressions are logical expressions --in a regimented language we list the logical expressions, and all of them will be among the ones for which we have no doubt that they are logical). Now, if each formal language is a model of how the meaning of a regimented language and a way the world might be affect the truth of the sentences of the
regimented language, and for each regimented language and a way the world might be we have a corresponding formal language that models it in the sense we just mentioned, we can say that a sentence is a logical truth iff it is 'true' in all formal languages that share the same system of forms as the regimented language. Of course, once we talk about formal languages it does not really matter if the property that the sentences have is that of being true or simply, say, that of being assigned the value 1. All that matters is that it is a property that can model the property of being true that the sentences of the regimented language do indeed have. For this it is enough that it be one of two properties that will be assigned to the sentences depending in the right way

16 Here is, very schematically, a justification of the bi-conditional:

Suppose that S is a first-order system of forms, and that the value that models (structures, interpretations) assign to each sentence of S is either 1 or 0, 1 being the value that all models assign to $\exists x x=x$. We can say that a model M corresponds to the combined effect of an interpretation (as a regimented language) of S, I, and a way the world could be, W, iff (1) holds (i) $\forall \alpha (M$ assigns 1 to $\alpha$ iff what $\alpha$ would say according to I is the case in W).

Analogously, we also say that an interpretation for S, I, and a way the world might be W correspond to a model M iff (i) holds. Let’s see that the two directions of the bi-conditional in the main text hold:

(a) Given some I and W let’s see that there is a corresponding M: Given I and W let A be the set $\{\alpha: \alpha \in S$ and $\alpha$ is true with respect to I and W$\}$; from A we can not deduce $\exists x x=x$ (otherwise, since the deductive rules are intuitively truth preserving, $\exists x x=x$ should be true according to I and W, but it is not); By completeness, it is not the case that $A=\exists x x=x$; so there is a model that assigns 1 to all the sentences in A; so M assigns 1 to exactly those sentences that are true with respect to I and W.

(b) Given some M, let’s see that there are some corresponding I and W: Let W the way the world actually is; let the domain of I be the individuals that are on the set-domain of M. Let each constant of S be a name of the object which is its value according to M; as for the meaning of a predicate symbol $P$, if B is the set that is the value of $P$ according to M, let the meaning of $P$ according to I to be such that $P$ would translate as is a member of C in a metalanguage where C was a name for B. Analogously for n-place predicates.

190
on the values that the expressions in the sentence have. (In fact, given that the sentences of a formal language do not really mean anything, probably it can not be said in a proper sense that they are true or false).

Now, given the last rendering of what it is for a sentence to be a logical truth, and given that which formal languages we have is determined by which models or structures we have, we can say that a sentence of a formal language is logically true iff it is true (has value 1, etc) in all models. This is, of course, the standard formulation.

Notice that if we took a formal language (or, equivalently, if we took the interpretation of a system of forms in a structure) just to model one possible meaning that the non-logical expressions can have (and not the combined effect of a possible meaning and a way the world might be), then that a sentence is true in all models would not, by itself, 'capture the idea', 'correspond to' or 'model the fact' that the sentence is a logical truth. Consider the following metaphor: there is a person in a room, there is a light behind him in the wall whose position can be changed to any point in that wall, the person projects a shadow on the floor of the room. How the shadow of the person is, and in particular some property of it like, say, that there are two points inside the shadow that are at least 2 yards apart, depends both on the position that the person adopts and on the location of the light. The position of the person's body is the meaning of a sentence, the location of the light in the wall is the way that the world is, the property of the shadow of having two points which are 2 yards apart corresponds to the sentence being true. If we want to explicate what it means that the shadow has the property just by virtue of the
position of some part of the person's body, say, the arms, we can not just say: it consists in the shadow having the property whatever the position of the other parts of the body is if we keep fixed the light in some specific location on the wall. Rather what we should have is: the shadow has the property just by virtue of the position of the arms iff the shadow would have that property whatever the position of the other parts of the person's body is and whatever the location of the light in the wall is.

To conclude, let's stress what is already indicated above about what we claim and what we do not claim regarding our 'third-way' view of the standard model theoretic analysis.

We are not claiming that when logicians apply the standard definitions in their daily practice they are actually viewing models in the way we contend is most appropriate to view them. Neither are we claiming that if we regard each model (or, equivalently, each formal language) as modeling a combination of one way the world might be and one meaning (as a regimented language) that certain system of forms can have, and then we use the standard model-theoretic analysis of the logical properties we are going to end up with results that are extensionally different from what we would get if we were regarding each model as modeling just one possible meaning of the language (while keeping the world fixed). There is no such extensional difference in, for instance, the case of standard first order languages.

What we do claim is that there is a difference at the conceptual level: if we view what we do in accordance to our 'third-way' approach then we can see why the standard model-
theoretic definitions of the concepts of logical truth and logical consequence are good analysis of the intuitive concepts, since they capture the essential features of those concepts --like, for instance, their modal character. If we view models in accordance to the interpretational perspective we can not justify at a conceptual level that the standard definition capture all the features of the intuitive notions that we want to capture. We can, of course, justify that even if we view models under the interpretational perspective, we get the appropriate extensions when applying the standard definitions. I contend, though, that we can see the standard definitions as providing more that just an extensionally correct method of determining which formulas are logically true, or which formulas follow logically from other certain formulas.

One question that would require further discussion but which we will not examine here is this: the standard model-theoretic analysis of the logical properties can be seen as having two parts. One the one hand, we have what is properly the conceptual analysis of the logical concept, where we say, for instance regarding the concept of logical truth, that a sentence is logically true if it is true in all models that correspond to one combination of a way the world might be and a meaning the sentences could have. Only with this first part we are not yet able to apply the analysis to any specific sentence. We need to know which all the models are that correspond to a combination of a way the world could be and a meaning the expressions could have (in order to see whether the sentence is 'true' in all of them). In the second part we throw in our substantive metaphysical assumptions about how the world could be, and what sets we actually have to model them, in order to determine which are all the models that should be considered when applying the analysis that we arrived at in the first part.