FREGE’S PUZZLE, COGNITIVE VALUE AND DIRECT REFERENCE

Problema de Frege, valor cognitivo e referência direta

Campinas

2015
FREGE'S PUZZLE, COGNITIVE VALUE AND DIRECT REFERENCE

Problema de Frege, valor cognitivo e referência direta

Dissertation presented to the Institute of Philosophy and Human Sciences of the University of Campinas in partial fulfilment of the requirements for the degree of Master in Philosophy.

Dissertação apresentada ao Instituto de Filosofia e Ciências Humanas da Universidade Estadual de Campinas como parte dos requisitos exigidos para obtenção do título de Mestre em Filosofia.

Orientador: PROF. DR. MARCO RUFFINO

Co-Orientador: PROF. DR. LUCAS ANGIONI

ESTE EXEMPLAR CORRESponde à
VERSÃO FINAL DA DISSERTAÇÃO
DEFENDIDA PELO ALUNO Filipe
Martone de Faria, E ORIENTADO
PELO PROF. DR. Marco Antonio Caron
Ruffino.

Campinas

2015
Martone, Filipe, 1988-
M367f  Frege's Puzzle, cognitive value and direct reference / Filipe Martone. – Campinas, SP : [s.n.], 2015.

Orientador: Marco Antonio Caron Ruffino.
Coorientador: Lucas Angioni.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Filosofia e Ciências Humanas.


Informações para Biblioteca Digital

Título em outro idioma: Problema de Frege, valor cognitivo e referência direta
Palavras-chave em inglês:
Languages, philosophy
Semantics
Epistemology
Área de concentração: Filosofia
Titulação: Mestre em Filosofia
Banca examinadora:
Marco Antonio Caron Ruffino [Orientador]
Ludovic Soutif
Emiliano Boccardi
Data de defesa: 28-08-2015
Programa de Pós-Graduação: Filosofia
A Comissão Julgadora dos trabalhos de Defesa de Dissertação de Mestrado, em sessão pública realizada em 28 de agosto de 2015, considerou o candidato FILIPE MARTONE DE FARIA aprovado.

Este exemplar corresponde à redação final da Dissertação defendida e aprovada pela Comissão Julgadora.

Prof. Dr. Marco Antônio Caron Ruffino

Prof. Dr. Ludovic Soutif

Prof. Dr. Emiliano Boccardi
For the three ladies who made me who I am today, Ana, Aura and Ada.
Acknowledgements

First and foremost I would like to thank my mother for her unconditional love, support and respect, without whom I would never be where I am now.

I also thank Gabriela, for always pushing me forward and for always saying the right (and harshest) things. And also for the jokes.

I am also grateful beyond measure to Lucas, Gregor, Leo and Gustavo for the many discussions, incentive and heartfelt friendship ever since I can remember.

Needless to say, I cannot thank my supervisor and friend Marco Ruffino enough for his inestimable help and tremendous encouragement, for being the incredible person and professional that he is, and of course for the countless bottles of cheap beer in the shadiest corners of Barão Geraldo. I am also deeply indebted to my co-supervisor and friend Lucas Angioni for his absolutely essential support in the earliest days of my studies.

I also would like to thank Nathan Salmon for being my sponsor during my time as a visiting scholar at UCSB, and for his valuable and extensive comments on the second chapter of this dissertation.

Many, many thanks to Emiliano Boccardi and Ludovic Soutif for reading my dissertation so thoroughly and for asking me the toughest questions. My work has benefited immensely from their insights and suggestions.

And last, but not least, many thanks to Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) (2013/22364-4), for funding the research that made this dissertation possible.
"Not at all, not at all, my dear sir! Let me see, I don’t think I know your name?"

"Yes, yes, my dear sir – and I do know your name, Mr. Bilbo Baggins. And you do know my name, though you don’t remember that I belong to it. I am Gandalf, and Gandalf means me!

J.R.R. Tolkien, The Hobbit: or There and Back Again
Abstract

Should we deal with Frege’s Puzzle on semantic grounds? Is the cognitive value of language an aspect of meaning? Frege himself assumed an affirmative answer to those questions, and so did many direct reference theorists, such as David Kaplan and John Perry. Even though they defend a theory that is anti-Fregean by nature, they share the view that it is semantics’ business to account for Frege’s Puzzle and cognitive value. There are two traditional ways to do so in referentialist semantics. One is via character and the other via reflexive content. My aim in this dissertation is to argue that both fail. To do that, I first examine what exactly Frege’s Puzzle is, and if what traditionally goes under the name of “Frege’s Puzzle” really corresponds to the puzzle that Frege himself formulated. I then examine how the solutions to the puzzle in terms of character and reflexive content are supposed to work for indexicals, where they are most appealing, and for proper names. I argue that there is no version of these solutions that is able to account for all the relevant phenomena. I conclude that, if this is the case, then we have serious reasons to suspect that Frege’s Puzzle should not be explained by semantics, and that cognitive value is not an aspect of meaning as it is often supposed.

Keywords: Frege’s Puzzle; cognitive value; reference; semantics; epistemology.
Resumo

Devemos lidar com o Problema de Frege dentro da semântica? É o valor cognitivo da linguagem um aspecto do significado? O próprio Frege assumiu uma resposta afirmativa a essas questões, assim como vários teóricos da referência direta, como David Kaplan e John Perry. Apesar de defenderem uma teoria semântica que é anti-fregeana por natureza, eles compartilham da concepção de que é tarefa da semântica resolver o Problema de Frege e explicar fenômenos de valor cognitivo. Há duas maneiras tradicionais de se fazer isso numa semântica referencialista. Uma é via caráter e outra é via conteúdo reflexivo. Meu objetivo nesta dissertação é argumentar que ambas falham. Para isso, primeiramente examino o que é exatamente o Problema de Frege, e se o que tradicionalmente é chamado de “Problema de Frege” na literatura corresponde ao que Frege tinha em mente. Depois disso, explico como supostamente funcionam as soluções ao Problema de Frege através do caráter e através do conteúdo reflexivo no caso dos indexicais – onde elas são mais plausíveis – e no caso de nomes próprios. Argumento que nenhuma versão dessas soluções é bem sucedida em explicar todos os fenômenos que devem ser explicados. Concluo então que, se esse é o caso, então temos boas razões para acreditar que o Problema de Frege não deve ser solucionado dentro da semântica, e que o valor cognitivo não é um aspecto do significado como comumente se supõe.

Palavras-chave: Problema de Frege; valor cognitivo; referência; semântica; epistemologia.
# Index

Introduction .................................................................................................................. 11

1. Frege’s Puzzle and cognitive value ........................................................................ 16
   1.1 Introducing the general puzzle ........................................................................ 16
   1.2 Frege’s version of Frege’s Puzzle .................................................................... 21
   1.3 Are the puzzles different after all? ................................................................. 40
   1.4 Concluding remarks ......................................................................................... 44

2. Indexicals ................................................................................................................ 45
   2.1 Pure indexicals and demonstratives ................................................................... 45
   2.2 Character and content ..................................................................................... 47
   2.3 Indexicals and Frege’s Puzzle ......................................................................... 49
   2.4 Counter examples ............................................................................................ 53
   2.5 Demonstrations to the rescue ........................................................................... 55
   2.6 Braun’s revision of Kaplan’s theory ................................................................. 65
   2.7 Kaplan’s revised theory and the solution (?) to Frege’s Puzzle ...................... 67
   2.8 A different solution ......................................................................................... 69
   2.9 Corazza and Dokic: acts of perception ............................................................ 73
   2.10 Final blow to the solution via character: pure indexicals .............................. 76
   2.11 The solution via reflexive content and concluding remarks ........................... 78

3. Proper Names .......................................................................................................... 81
   3.1 The character of names ................................................................................... 81
   3.2 Names and reflexive content .......................................................................... 89

4. Conclusion ............................................................................................................... 93

5. References ............................................................................................................ 95
Introduction

It is very common and intuitive to believe that language and thought bear an extremely intimate relation. Indeed, we frequently have the idea that language is some kind of vehicle for thought, a linguistic clothing that we employ in order to externalize the content of our minds. After all, we seemingly think via language, and when we want to convey our thoughts we simply voice out loud what is already in our heads. In this picture, language is somewhat a mirror or reflection of thought. Frege was no exception to this view. For him, language and thought are indeed closely connected: thoughts are expressed in the form of words and their parts correspond to the parts of the sentences used to express them1.

Despite being best known for his contribution to the philosophy of language, his primary philosophical interest was in thought, and not in language per se2. He explicitly says this in Frege (1956). However, thoughts are immaterial things. He could not manipulate them or look at them under a microscope for examination. The only way to access them, Frege believed, is through their linguistic embodiment. So, he had to turn to language in order to study thought; that is, he had to engage in semantics. Given his goal, it is clear that semantics does not have much importance in itself. Since his main philosophical concern was the thought, not language, an investigation of the semantic properties of linguistic expressions is relevant only if it captures and explains the functioning of thought. Thus, a semantic theory that does not account for the cognitive dimension of language is completely misguided. In other words, he believed that an account of the cognitive aspects of language must not be a secondary interest or a byproduct of a semantic theory, but rather its main concern. Semantics and epistemology, therefore, should be inextricably associated3.

The influence of this idea can hardly be overstated. With this conception, Frege established a model of semantics that would become orthodox during the 20th century: any

2 It is important to stress the fact that Frege was not interested in the psychological process of thinking – which perhaps should concern psychologists – but rather in the “eternal structure of thought”, as Burge (2005) points out. What this means is that his interest was in the form of rational thinking, in the laws of logic, assertion, judgment and inference, and not in how our mental acts come about.
3 As we will see in chapter 1, this interpretation of Frege is overtly simplified. For the purposes of this introduction, however, it is adequate enough.
semantic theory must not only explain the relation between words and world, but also the
relation between language and thought. In this Fregean approach to semantics\(^4\), questions
about the cognitive value (or cognitive significance) of language should be among the
primary concerns of semanticists. This is why Frege’s famous puzzle of identity (which is
not restricted to identity) was and still is a much debated topic in the philosophy of
language and mind. The puzzle can be roughly stated as follows: how can sentences of the
form \(a=a\) be non-informative and trivial and sentences of the form \(a=b\) be informative if \(a\)
and \(b\) are coreferential\(^5\)? It seems evident that the same speaker can be bored by sentences
of the first form but be rather shocked when presented to the sentences of the second form.
For example, a speaker may find the sentence “Jack the Ripper is Jack the Ripper” idiotic,
although she would (presumably) be terrified to be told that “Jack the Ripper is Will”,
where “Will” is the name of her husband (for a less dramatic case, think of Aunt May
being told that “Spider-Man is Peter Parker”). How can the mere substitution of
coreferential terms affect the cognitive value of those sentences? If semantics must be
epistemically sensitive as Frege believed, then it must provide an answer to this question in
terms of the semantic properties of the expressions involved. In other words, cognitive
value must be explained in terms of the (literal) meaning of these expressions.

This is what became known in the literature as Frege’s Puzzle. Frege originally
stated it using schemas, or sentences involving proper names and definite descriptions, but
it is a much more general phenomenon. Given Frege’s philosophical interests and his view
of semantics as a means to study thought, it is not surprising that this puzzle is seen by him
as perhaps the central problem in the philosophy of language.

What is striking, however, is that even with the widespread reaction against
Frege that started in the late sixties, many philosophers also felt the need to deal with this
puzzle on semantic grounds. In other words, many philosophers that eschewed Fregean
semantics worked and still work under the Fregean assumption that it is semantics’
business to account for the problem of cognitive value, i.e., to deal with Frege’s Puzzle in
its various forms. Among these philosophers were some paradigms of direct reference such
as David Kaplan and John Perry. They too felt the pull of the puzzle and devoted a

\(^4\) I will use “Fregean model of semantics” and “Fregean approach to semantics” as synonyms, meaning the
idea that explaining cognitive value is a criterion of adequacy for semantic theories.

\(^5\) Note that it is not entirely obvious what \(a=a\) and \(a=b\) are supposed to represent. I will deal with this
problem in the first chapter.
significant part of their work to solve it. In sum, a great number of philosophers abandoned Fregean semantics after the sixties and adopted the so-called direct reference theory; however, they did not abandon the Fregean approach, that is, the idea that semantics is inseparable from epistemology and that it must contemplate the cognitive dimension of language. They tried to accommodate anti-Fregean semantics in the Fregean model.

Obviously, if these philosophers abandoned Frege, the famous and elegant solution for the puzzle that he offered in Über Sinn und Bedeutung was of no use to them. They had to employ only the tools available to the new semantic theory that emerged, which is Millian in spirit. The big problem is that Frege himself rejected all kinds of Millian semantics precisely because they are incapable of explaining Frege’s Puzzle. The reason is the following. Mill maintained that the semantic value or content of a name is exhausted by its reference. There is no other content associated with the name that has any kind of semantic relevance. Of course, a name can evoke images, descriptions or any sort of qualitative content for a speaker or even for a whole community, but this associated content does not perform any semantic function whatsoever.

According to Frege, however, if the sole semantic value of a name is its reference, then the sentences of the form $a=a$ and $a=b$ would express the same thing: the trivial proposition that $a=a$. It would not matter if we substitute $x$ in $a=x$ for $a$ or $b$; their semantic value would be exactly the same entity, and hence the same content would be expressed either way. For Frege this is absurd, for it is possible, as noted in the examples above, that a competent speaker accepts one but sincerely denies the other. How can this be the case if both sentences express precisely the same thing? If they say the same thing, they ought to have exactly the same cognitive value. Different attitudes should be caused by different contents. Millian semantics has the consequence that a speaker may not fully capture or fail to recognize a proposition expressed by sentences that he fully understands or, worse yet, that he himself expresses. In other terms, it generates a split between what is expressed by a sentence and what is captured by the speaker’s mind. In Millian semantics, a difference in cognitive value does not necessarily reflect an objective semantic difference. The so-desired unity between semantics and epistemology seems to fall apart. It appears that this kind of referentialist semantics is not capable of dealing with the cognitive phenomena that were so important to Frege.
The same problem plagues direct reference. If the Millian spirit returns, the difficulties in explaining cognitive value on semantic grounds return as well. Direct reference theorists argue that not only names have their semantic values exhausted by their referents, but also indexicals, natural kind terms and even some uses of definite descriptions. As we saw, Millian semantics seems to generate a disagreement between what is expressed by sentences and the cognitive life of the speakers: even if it is possible to have distinct attitudes towards the sentences $a=a$ and $a=b$, they nevertheless express exactly the same (singular) proposition. Perhaps the most emblematic example of this split between semantics and cognition in direct reference is the poor Pierre, who in virtue of accepting both the English sentence “London is not pretty” and the French sentence “Londres est jolie”, believes that London is pretty and not pretty at the same time, for the semantic value of both “London” and “Londres” is just London and the propositions they express are identical. The Fregean marriage between semantics and epistemology seems to have gone through a terrible divorce.

How can direct reference theorists account for Frege’s Puzzle if Millian semantics seems ill suited to this purpose by its very nature? Despite initial appearances, the new and more sophisticated Millians have good resources at their disposal for dealing with cognitive value. There are two main ways to pursue the solution to the puzzle on semantic grounds: one via character, most prominently defended by Kaplan, and other via reflexive content, most prominently defended by Perry. These solutions became somewhat standard, especially the solution in terms of character in the case of the indexical version of the puzzle. Direct reference theorists, even being Millian heirs, managed to find their way out of Frege’s Puzzle while still conforming to the Fregean approach to semantics. The problem is, I think, that these solutions do not work.

My purpose in this dissertation is twofold. I want first to investigate what exactly Frege’s Puzzle is, and then argue that the traditional direct reference attempts to solve it on semantic grounds fail. If I am right and they do fail, this lends extra weight to Wettstein’s (1986) famous challenge to the Fregean criterion of adequacy for semantics: if the most promising referentialist resources available to deal with the puzzle are unable to do it adequately, then we have very good reasons to suspect that explaining cognitive value should not be the semanticist’s business in the first place. In other terms, if character and

---

6 This point about definite descriptions is not uncontroversial, however.
reflexive content cannot explain all relevant phenomena of cognitive value successfully, then it seems that cognitive value is not an aspect of meaning after all, and we might consider abandoning the Fregean approach to semantics.

In the first chapter, I first present the general puzzle that commonly goes under the name of “Frege’s Puzzle” in the literature. I then investigate if this general puzzle really corresponds to Frege’s own version of the puzzle. To do that, I try to dispel some misunderstandings that arise in relation to the way Frege poses the puzzle in Über Sinn und Bedeutung. His formulation in this work is rather unfortunate on several grounds. A better understanding of what Frege had in mind with the puzzle may help to understand what is essential to its formulation, and what exactly is the problem he believed to be solving. This is especially important given Glezakos’ (2009) charges against Frege’s formulation of the puzzle: she argues that he cannot pose it in a theory neutral manner, so we should not be puzzled by it at all.

In the second chapter, I deal with the indexical version of Frege’s Puzzle. I introduce Kaplan’s theory of indexicals and explain how it is supposed to solve the puzzle in terms of character. I argue that neither his nor some modified versions of character can account for cognitive value successfully. I also present Perry’s theory of reflexive content and how it manages to apparently solve the puzzle for indexical expressions. I then show why it fails.

In the third chapter I present the puzzle as arising for proper names. I argue that any attempt to solve it in terms of character requires a major deviation from Kaplan’s original theory and from standard referentialism, and that these attempts ultimately fail as well. I then briefly explain how reflexive content could perhaps account for the cognitive value of names, but not on semantic grounds. If this is right, then no aspect of the meaning of directly referential expressions is able to solve Frege’s Puzzle, and thus we have compelling reasons to believe that cognitive value is not a feature of meaning.
1. Frege’s Puzzle and cognitive value

1.1 Introducing the general puzzle

Suppose you are a New Yorker accustomed to seeing Spider-Man swinging by the buildings in Manhattan. You often read about his incredible doings in the newspapers and see him on TV every other day. You admire his strength and righteousness. You also happen to live right next to a nerdy and unremarkable kid called Peter Parker who is always being bullied by his schoolmates. You even have given him a ride to the library once or twice. Imagine your surprise if you encounter the following newspaper headline on your way to work: “Spider-Man is Peter Parker!” That very day, you return home and find so many press vehicles and journalists on your street that you are unable to count them. They want to interview you, hoping to get a glimpse at Parker’s life, and you are harassed for several days. You learn that his poor aunt was so shocked by this news that she suffered a heart attack and is now hospitalized. With Parker as inspiration, an anti-bullying organization called “Spidey’s Friends” is founded and nerdy kids everywhere start standing up against their bullies. The mayor even names a high school after Peter Parker. After the bombastic headline, life in New York City is dramatically (and perhaps forever) changed.

Now, imagine that instead of “Spider-Man is Peter Parker!” the headline was “Spider-Man is Spider-Man!” or “Peter Parker is Peter Parker!” Would those headlines cause such generalized commotion and such profound changes in the city life? The answer seems obvious: no. The life of New Yorkers would go on just as the same. At most, the responsible editor would be fired for wasting the front page with such a boring triviality. There would be no harassment, no heart attack and unfortunately no anti-bullying organization. These headlines would be just silly.

But how can we explain such drastic difference in the outcome of those headlines? The names “Spider-Man” and “Peter Parker” are names of the same individual. If this the case, then the sentences “Peter Parker is Peter Parker” and “Spider-Man is Peter Parker” should say precisely the same thing about that individual, i.e., that he is one and the same. But it seems obvious that they do not. How could they? The first sentence is completely uninteresting, while the second is capable of enormously affecting the behavior
of a whole community. How can the mere substitution of one name for another so greatly modify the impact that a sentence has on the speakers’ cognition if those names refer to the same individual? How come one is trivial and the other informative?

Examples like this one abound. Consider the following set $S$ of sentences:

a. Hesperus is Hesperus
b. Cicero is Cicero
c. Batman is Batman
d. Walter White is Walter White
e. Darth Vader is Darth Vader
f. Freddie Mercury is Freddie Mercury

All sentences in set $S$ seem pretty trivial and uninteresting. No one in her right mind would assert one of them literally and in all seriousness. Now, compare set $S$ with set $S'$:

i. Hesperus is Phosphorus
ii. Cicero is Tully
iii. Batman is Bruce Wayne
iv. Walter White is Heisenberg
v. Darth Vader is Anakin Skywalker
vi. Freddie Mercury is Farrokh Bulsara

The difference is striking. Sentence (i) expresses an important astronomical discovery, while sentence (a) does not. Sentence (iv) expresses what Albuquerque police spent the five seasons of *Breaking Bad* trying to figure out, while everybody knew that sentence (d) was true since the beginning of the story. Sentence (c) is silly to everybody in Gotham City, but sentence (iii) would cause a fuzz similar to the Peter Parker/Spider-Man case above. Sentence (v) expresses one of the most dramatic revelations in movie history, but not sentence (e). Even though the names occurring in those pairs of sentences refer to the same thing, these sentences have distinct impacts on the cognitive life of speakers: Set

---

7 I present the puzzle in terms of sentences, but the point could very well be made in terms of sentence tokens or utterances.

8 I am obviously not using the term “express” here in the sense of expressing a proposition. This term must be understood in the loosest sense possible, without any theoretical commitments. I just mean that they communicate or convey such information.
S contains only trivialities, while sentences in set $S'$ convey important information that could affect the speakers’ behavior in often drastic ways. A speaker may even accept unquestionably all sentences in set $S$ and at the same time doubt or vehemently deny every single sentence in set $S'$ without being considered crazy or irrational. Let us call the properties that a sentence has that are capable of affecting the speaker’s attitudes towards it its cognitive value. Thus, the triviality of sentences in set $S$ and the informativeness of sentences in set $S'$ are aspects of their cognitive value.

Identity statements are not the only sentences that may differ in cognitive value. Sentences predicating properties of an object also may exhibit the same sort of cognitive difference. Consider these two sentences:

a. Freddie Mercury is one of the greatest rock singers of all time
b. Farrokh Bulsara is one of the greatest rock singers of all time

I may strongly believe sentence (a), but start a violent argument with you if you assert sentence (b). How can this be possible if we are talking about the same person and predicating exactly the same thing of him? How can my attitude towards the first be different than my attitude towards the second? It could be argued that our disagreement arises because I am not competent with the name “Farrokh Bulsara”. But this is not the case. Suppose he is my brother, but he managed to hide his artistic career from me all his life, so I am quite dubious of his singing abilities. I am then as competent as possible with both “Farrokh Bulsara” and “Freddie Mercury”. Yet, I believe that sentence (b) is false, while believing that sentence (a) is the most obvious truth ever known to man. A difference in cognitive value, then, does not arise by lack of competence with the relevant proper names.

It gets worse. Proper names are not to blame: definite descriptions, indexicals and kind terms also affect the cognitive value of the sentences in which they occur. Consider these sentences:

a. The morning star is the morning star
b. The morning star is the evening star
c. The morning star is bright
d. The evening star is bright
a. I am me [Bruce Wayne pointing at himself]
b. I am he [Bruce Wayne pointing at some footage of Batman on TV]
c. I fight criminals at night [said by Bruce Wayne in his Batman suit]
d. He fights criminals at night [said by Alfred pointing at Bruce Wayne]

a. Furze is furze
b. Furze is gorse
c. Furze has thorns
d. Gorse has thorns

It seems clear that the cognitive value of those sentences is different, even though the expressions flanking the copula refer to the same thing. Sentences (a) are trivial, while sentences (b) are informative. A speaker can sincerely accept (c) sentences and at the same time sincerely reject (d) sentences, and vice-versa. Remember that this difference is not due to any kind of linguistic incompetence: the speaker masters completely the use of every referring expression contained in those sentences, an even so their cognitive values are distinct.

The problem in its most general form, then, is this: how can the substitution of coreferential expressions (names, indexicals, definite descriptions, etc.) affect the cognitive value of a sentence? What accounts for the epistemic properties that sentences exhibit, such as triviality and informativeness, and their potential to affect the behavior of speakers? Is cognitive value an aspect of the meaning of those sentences or of something else? These questions are extremely important. They concern the relationship between meaning, reference, logic and the epistemic dimension of language, as well as the question of how we acquire and process information by linguistic means. This problem is what became known in the literature as Frege’s Puzzle, and it dominated much of the philosophy of language ever since it was introduced.⁹

⁹ It is worth noting that I presented the puzzle here only on behavioral grounds, i.e., considering only the way speakers produce and react to certain sentences (or their utterances). I believe this is the best way to set up the puzzle, because it does not involve (as far as I am aware) any problematic theoretical assumptions. In other words, setting up the problem based on the observed behavior associated with certain sentences (or their utterances) avoids any prior commitments regarding the determination of their syntactical and logical form, their truth-conditional content and the relation of those features with their epistemic profiles. The puzzle, as presented here, rests only on the general fact that we often refer to and think about the same object
Yet, it is not obvious that this puzzle, as presented above, corresponds exactly to the puzzle that Frege himself was concerned with. This puzzle’s precise nature is a matter of debate, given that Frege’s formulation involves many more subtleties than have been presented so far. There are several interpretations on the market (e.g. Salmon (1986); Salmon (1992); Mendelsohn (2005); Dickie (2008)) which are far from reaching consensus. This is quite remarkable, since much of the philosophy of language revolved around a puzzle whose precise features are subject to significant disagreement (Dickie (2008), for example, argues that Frege had not one, but two puzzles that we have inadvertently lumped together).

Especially worrisome, however, are the interpretations contained in Glezakos (2009) and in Taschek (1992). Glezakos claimed that the puzzle does not even get off the ground, because it presupposes, in order to be posed, the very notion that Frege is aiming to introduce (Sinn). In other words, she argues that Frege’s puzzle is unavoidably question begging, so there is nothing to be puzzled about. If this is true, then the purported solutions I aimed to discuss here are solutions to a pseudo problem, and there would be no point in examining how well they fared in dealing with it. Taschek, on the other hand, claims that there is a puzzle all right, but direct reference theorists have completely missed its point. The original puzzle, he says, has to do with the nature of logic and its role in our linguistic and epistemic life. The solutions these theorists offered, then, do not engage with the real issue that needs to be addressed.

So, in order to be able to assess the plausibility of these striking claims (which, if true, would probably render this dissertation pointless), I will first present my own take on the original puzzle. After all, how can I examine solutions to a problem without knowing clearly what the problem is (if there is a problem), or at least which version of the problem was addressed by them? I will not discuss all interpretations that are out there. I will only mention them when I see fit. This of course does not mean that I totally disagree with them or that my interpretation is completely novel; it only means that, to me, those interpretations are not wholly convincing or comprehensive enough, so I will make an effort to put together everything I find relevant.

through referential expressions without realizing we are doing so, and there is often a change in our epistemic state when we apprehend certain identity statements and not others. As we will see, Frege’s version of the puzzle is somewhat different. However, this puzzle still is a puzzle, and even if it is not exactly what Frege had in mind, it must be accounted for.
1.2 Frege’s version of Frege’s Puzzle

The locus classicus of Frege’s presentation of the puzzle is the opening passage of Über Sinn und Bedeutung (henceforth SuB):

Equality gives rise to challenging questions which are not altogether easy to answer. Is it a relation? A relation between objects, or between names or signs of objects? In my Begriffsschrift I assumed the latter. The reasons which seem to favor this are the following: \( a = a \) and \( a = b \) are obviously statements of differing cognitive value (Erkenntniswerte); \( a = a \) holds \textit{a priori} and, according to Kant, is to be labelled analytic, while statements of the form \( a = b \) often contain very valuable extensions of our knowledge and cannot be always established \textit{a priori}.

(Frege, 1960a, p. 56)

The question he asks, then, is how sentences of these two different forms can have distinct epistemic profiles if both names flanking the identity sign refer precisely to the same object. In other words, Frege wants to know how it is possible for sentences of the form \( a = b \) (with \( a \) and \( b \) coreferential) to contain “valuable extensions of our knowledge” if they require, in order to be true, the exact same conditions as sentences of the form \( a = a \).

As we can see, Frege’s formulation of the puzzle involves certain notions whose precise characterization is not entirely obvious. What exactly are the schemas \( a = a \) and \( a = b \) supposed to represent? What is his conception of analyticity and of \textit{a prioricity}? What does he mean by “valuable extensions of our knowledge” and what is its relation to \textit{Erkenntniswerte} (cognitive value)? Answering those questions and interpreting the passage above is a notoriously tricky business, as the many different interpretations available can attest. In part, this is Frege’s own fault. His formulation of the puzzle in this passage is unfortunate in several ways. It involves, to my view, many unnecessary complications that cloud the heart of the matter and that can give rise to mistaken interpretations. If we remove at least some of the noise from this passage, we might arrive at a better appreciation of what is really at stake. In what follows, I will try to do so. I claim that this formulation is infelicitous in mainly three aspects: (1) setting up the puzzle in terms of identities; (2) invoking the notions of analyticity and \textit{a prioricity}; and (3) setting up the puzzle in terms of the schemas \( a = a \) and \( a = b \) without being explicit about what they represent.
The first and most obvious aspect in which this way of posing the puzzle is unfortunate is its formulation in terms of identity statements. As it has now been widely recognized (e.g. Salmon (1986), Mendelsohn (2005), Boccardi (2014)), identity statements are pretty much inessential to the puzzle. They are just where the problem “happens to show its virulence” (Mendelsohn, 2005, p. 30). Frege apparently was not aware of this fact when he wrote his *Begriffsschrift* (henceforth *BS*), when the puzzle first showed up, but he certainly was by the time of *SuB*. Evidence of this is that in *Funktion und Begriff* (henceforth *FuB*), written about a year before *SuB*, Frege presents the puzzle by means of the following sentences, which are evidently not identity statements:

(a) The Evening Star is a planet with a shorter period of revolution than the Earth
(b) The Morning Star is a planet with a shorter period of revolution than the Earth

Frege notices that the proper names\(^{10}\) “the Morning Star” and “the Evening Star” contained in those sentences refer to the same heavenly body – Venus – and yet the thoughts expressed by (a) and (b) are different, “for somebody who does not know that the Morning Star is the Evening Star might regard one as true and the other as false” (Frege, 1960a, p. 29). He also explicitly claims a bit earlier in the same passage that “from identity of reference there does not follow identity of the thought [expressed]” (Frege, 1960a, p. 29). He then goes on to mention his sense/reference distinction as the explanation of this phenomenon, exactly as he does in *SuB*. This passage from *FuB* shows that he is perfectly aware that substituting coreferential expressions (proper names, for that matter) may affect the cognitive value of any sentence that contains them, not only identity statements.

Overall, this is a far cleaner and less contentious formulation of the puzzle, for it involves real sentences, not schemas, and has fewer problematic assumptions than the formulation contained in *SuB*\(^{11}\). Of course, the notion of “thought” here is far from being trivial, but Frege does not need to explain it at this point. To set up the puzzle it is enough to appeal to the reader’s intuition regarding the relevant cognitive difference that arises by

---

\(^{10}\) In his sense of proper name (*Eigennname*), which encompasses not only proper names, but also definite descriptions.

\(^{11}\) However, as Boccardi (2014) notes, this formulation is subject to the same criticism that Glezakos’ (2009) put forward against the formulation of *SuB*, for it also apparently relies on some criterion of name individuation. As we will see, her criticism misses the real point of the puzzle.
substituting one coreferential expression for the other in sentences (a) and (b). It is quite mysterious why he chose a much more complicated manner of presenting the puzzle in *SuB* given that he had a much simpler one at his disposal.

Whatever his reasons are, the fact is that posing the puzzle only in terms of identity statements is more confusing than helpful. The first and obvious problem is that this may lead to a mistaken appreciation of the scope of the puzzle. By reading the opening paragraph of *SuB*, we might easily believe that the problem is restricted to identities (as Frege himself did in *BS*) and fail to see that it appears for other types of sentences as well.

The second problem is that his exposition associates rather obscurely the *metaphysical* problem of identity with the *epistemic* problem of identity. The metaphysical problem concerns the nature of identity relations themselves, while the epistemic problem has to do with the epistemic properties of statements expressing those relations. Frege apparently means to investigate the epistemic problem by raising the metaphysical one, but he is not very clear about how exactly they are related. Moreover, after introducing his distinction between sense and reference as a solution to the epistemic problem, he remains silent about how it is also supposed to solve the metaphysical one. Because this problem apparently was his main motivation for posing the puzzle in *SuB*, this lack of explanation is quite frustrating. In conclusion, then, given that identity is inessential to the puzzle, raising metaphysical issues about the identity relation in this context only obscures the matter.

A second aspect in which the opening paragraph of *SuB* is unfortunate is in its use of the notions of analyticity and *a prioricity* to characterize the relevant differences between sentences of the form *a=a* and *a=b*. Frege’s way of putting things easily suggests the following (false) opposition: because he first claims that sentences of the form *a=a* are always *a priori* and analytic and then goes on to contrast them with sentences of the form *a=b*, it is natural to assume that the latter are *a posteriori* and synthetic (since these are the opposite properties of *a prioricity* and analyticity) and that they express empirical claims. This last point is also suggested by the two examples he gives immediately after the quoted passage, both drawn from astronomy. Frege also claims that *a=b* sentences (but not *a=a*

---

13 Many different interpretations were offered over the years about how his solution is supposed to apply to the metaphysical problem. Thau and Caplan (2001), for instance, even argue for the (very) unorthodox claim that Frege never abandoned the metalinguistic view of identity that was first proposed in his *BS*. See Mendelsohn (2005), chap. 4 for a nice survey of the different interpretations of Frege’s change of mind that have been proposed.
sentences) are capable of extending our knowledge, a property which is most straightforwardly attributed to synthetic and a posteriori sentences, not to analytic and a priori ones. This reading, however, is way off the mark.

To be fair with Frege, he does say that the truth of \( a=b \) sentences cannot always be established a priori, so it follows that their truth can be established a priori at least sometimes. This observation clearly undermines the interpretation that \( a=b \) sentences are a posteriori and express empirical claims. At most, we can conclude that they can be a priori or a posteriori, and therefore do not necessarily express empirical claims. The examples he gives to illustrate \( a=b \) sentences, then, are somewhat infelicitous. If this is the case, then it is not very clear what the relevant contrast between \( a=a \) and \( a=b \) sentences is supposed to be. Being not-always-a-priori and not-always-empirical do not tell us much about the distinguishing features of \( a=b \) sentences, about what really makes them different from \( a=a \) sentences.

It seems that the only alternative is to say that the distinguishing feature of \( a=b \) sentences is that they are synthetic. This would neatly explain their potential, as opposed to analytic sentences, for extending our knowledge. After all, synthetic statements are, in their most intuitive conception, the ones that add something to our knowledge and not merely elaborate on the concepts we already possess. Besides, after Kant, the idea of synthetic a priori statements was already on the market, so it would be perfectly plausible to interpret Frege as holding that \( a=b \) sentences are synthetic, and that some of them can be known a priori and some can be known a posteriori. But why does he not explicitly say this in the text? Why does he say only that \( a=a \) sentences are always analytic, but not that \( a=b \) sentences are always synthetic, and that for this reason contain “valuable extensions of our knowledge”? It certainly would have made things much clearer.

He does not say that for a very simple reason: \( a=b \) sentences are not necessarily synthetic either. It is true that Frege’s most famous examples of sentences of the form \( a=b \) come from the empirical sciences (e.g. his Hesperus and Phosphorus example) and from geometry, whose truths, as Kant, Frege thought were synthetic.\(^{14}\) However, and this is important, he believed that some statements from arithmetic can also be of the form \( a=b \), where \( a \) and \( b \) refer to the same number but as resulting from distinct

\(^{14}\) See Frege (1960a), p. 11, and Frege (1960a), p. 57 for his famous examples from geometry. For the claim that geometry deals with synthetic truths, see Frege (1960b), p. 101-2.
operations (Ruffino, 2014, p. 36). As it is well known, Frege’s agreement with Kant regarding the existence of synthetic a priori truths is restricted to geometry; arithmetic, for Frege, is definitely analytic, since it is derived from the laws of logic. It follows, then, that some sentences, namely those from arithmetic, can be of the form \( a=b \) and be analytic nevertheless.

So, we end up with the following picture: \( a=a \) sentences are always analytic and a priori, while \( a=b \) sentences can have any of the three profiles –analytic and a priori, synthetic and a priori, or synthetic and a posteriori. As it seems clear, characterizing the two types of sentence forms in terms of these properties does not yield a sharp distinguishing criterion for them. Thus, if there is a real and relevant difference between sentences of the form \( a=a \) and \( a=b \), then there must be some property that one has that the other has not, which would explain why \( a=a \) sentences are always a priori and analytic while \( a=b \) sentences are not. What could it be?

In the quoted passage from Sub, the property that seems to be predicated only of \( a=b \) sentences, and hence the only one that really seems to distinguish them from \( a=a \) sentences, is being capable of extending our knowledge, whatever this means exactly. If the foregoing discussion is right, this property is not explained in terms of the sentence being a posteriori or synthetic, as it could be initially assumed, because analytic and a priori sentences can also be of the form \( a=b \) and thus are also able to extend our knowledge. Hence, Frege must have had something very different in mind about what extension of knowledge amounts to.

All this talk of analyticity and a prioricity in the opening passages of Sub, then, is completely dispensable. These notions are obviously not essential to the phenomenon that Frege is really interested in and do not mark any relevant difference between \( a=a \) and \( a=b \) sentences. In fact, they are more confusing than clarifying. What really matters for the puzzle in Sub is that \( a=a \) sentences are never capable of extending our knowledge, while \( a=b \) sentences are, regardless of them being analytic, synthetic, a priori or a posteriori. And, most importantly, \( a=b \) sentences extend our knowledge even referring to the same object twice and having the same truth conditions as \( a=a \) sentences. So, the question that Frege wants to answer in Sub, purged of all that irrelevant stuff, becomes this: what is it about sentences of the form \( a=b \) that makes them capable of extending our knowledge? In
what follows, for ease of exposition, I will call the property of extending our knowledge *

*informativeness*, and its contradictory *triviality*.

This brings us to the final and most misleading aspect of Frege’s formulation of the puzzle in *Sub*. Frege poses the puzzle in terms of the schemas $a=a$ and $a=b$ without really saying what those schemas are supposed to represent. He gives some examples and hopes that we get the point he is making, but unfortunately things are not so transparent. Without a clear grasp of what those schemas stand for, then, we cannot understand what he means by the property of informativeness and we risk misunderstanding what the puzzle is really about. In the next pages, I will present the most obvious interpretation of the schemas and show how Glezakos’ criticism applies to it. I will then argue that there is a different interpretation that is not subject to her criticisms and which is, to my view, much closer to what Frege really had in mind with the puzzle.

The most obvious reading of the passage in *Sub* is this. The schemas $a=a$ and $a=b$ are abstractions that represent the *general form* of identity statements, and the symbols flanking the identity sign represent *names*. Hence, a sentence will have the form $a=a$ when *the same name* occurs twice and will have the form $a=b$ when *two names* occur once. This seems to capture the relevant difference between sentences such as “Hesperus is Hesperus” and “Hesperus is Phosphorus”, or between “Peter Parker is Peter Parker” and “Peter Parker is Spider-Man”: the first sentence of the pairs have *only one name* occurring twice, and hence is of the form $a=a$, whereas the second sentence of the pairs involves *two names*, and hence is of the form $a=b$. After all, the names occurring in those sentences are clearly syntactically different, a fact that is captured by representing them with different symbols in the schemas. It appears obvious, then, that there is a significant difference between sentences of the form $a=a$ and $a=b$: the second is informative while the first is trivial, and the puzzle is to explain why this is the case if both have the same truth conditions.

So far, so good. The problem arises when we consider sentences where apparently only one name occurs, and yet there seems to be the same kind of cognitive difference that Frege is puzzled about. The most famous example of this situation is Kripke’s Pierre. Suppose that Pierre is a fan of the pianist called “Paderewski”. Suppose also that he disagrees vehemently with the policies of a certain polish politician called “Paderewski”. Unbeknownst to him, they are the same person. It seems that the sentence
“Paderewski is Paderewski” is informative to him even though there is only one name occurring twice. Or is it?

This is far from obvious. Should we say that the sentence “Paderewski is Paderewski” is of the form $a=a$ or $a=b$? Is there a single name occurring twice, as it seems, or two names? This puts Frege on the horns of a dilemma: either “Paderewski is Paderewski” is of the form $a=a$ and thus trivial, contrary to what he believes, or it is of the form $a=b$ and informative. Clearly, Frege must grab the second horn of the dilemma in order to be something to be puzzled about. But to do so, he must explain why the two occurrences of “Paderewski” in that sentence are occurrences of distinct names despite being tokens of the same syntactic type, and he must do this without presupposing the very notion he is aiming to introduce to solve the puzzle, namely, sense.

Here is where Glezakos’ criticism comes in. She claims that the only such theory neutral criterion available to Frege is contained in his famous footnote B from SuB: two occurrences of a name are occurrences of the same name iff they are two tokens of the same syntactic type and share the same referent (Glezakos, 2009, p. 204). Under this criterion, “Paderewski is Paderewski” would be of the form $a=a$, since the two occurrences of “Paderewski” have the same referent and are clearly two tokens of the same syntactic type. However, this sentence would be informative to Pierre despite being of the form $a=a$, as much as any sentence of the form $a=b$. Thus, there would be no relevant cognitive difference between $a=a$ sentences and $a=b$ sentences as Frege assumed. The puzzle is then dissolved.

She also claims that Frege might have proposed another criterion for name individuation in Der Gedanke. However, according to her, this criterion necessarily presupposes the notion of sense: two occurrences of a name are occurrences of the same name iff they have the same sense (Glezakos, 2009, p. 206). Of course, this criterion also dissolves the puzzle, for it assumes the very notion that is supposed to solve the puzzle. In other words, this way of individuating names is question begging; if we are not already attracted to the notion of sense on independent grounds, there is nothing to be puzzled about.

Hence, under the above interpretation of what the schemas $a=a$ and $a=b$ represent in SuB, either the puzzle disappears or it is circular. In either case, we might very
well dispense with it. However, I believe that this way of reading the opening passage of *SuB* does not do much justice to Frege and, more importantly, is not necessary. I think there is another possible interpretation of what he had in mind with the schemas $a=a$ and $a=b$ that avoids Glezakos’ objections and that, to my view, gets closer to the heart of the puzzle. In order to present this interpretation, however, it is necessary first to go over some important aspects of Frege’s philosophical project.

Frege’s primary goal was to study thought. Not thought understood as the psychological processes of cognition, as mental occurrences which happen in time and in some particular brain. He wanted to study thought in an abstract sense, as that “for which the question of truth arises” (Frege, 1956, p. 292). Thoughts, for Frege, are the sort of things that can be true or false, that are expressed by language and are what we apprehend when we think. They are not generated or created by us: we merely grasp them. Thus, thoughts in Frege’s sense are objective, immaterial and eternal things which are the bearers of truth and falsity and can be apprehended and communicated by different persons at different times.\(^{15}\)

The discipline that studies thoughts understood in this sense is *logic*. And the goal of logic, Frege claims, is to find what are the most general laws that describe the relations and connections between thoughts, in much the same way as physics describes the laws of nature and biology describes the laws of life. In other words, Frege believed that thoughts are related to each other in such a way that some thoughts *follow naturally* from others, i.e., that thoughts enter into deductive relations with one another following a *natural order*.\(^{16}\) The task of the logician, then, is precisely to uncover the rules that govern this natural ordering, i.e., to describe the deductive relations that thoughts can have with each other, thus explaining why some thoughts are logical consequences of others whereas others are not. And this can be done regardless of the subject matter at hand: logic describes the laws of thought not as applied to this or that specific domain, but of thought *qua* thought, as something which is common to *all* possible domains. Logic, then, is absolutely general. Moreover, in finding the laws that describe the natural ordering of thoughts, the logician explains why moving from one thought to another in a chain of

---

\(^{15}\) In what follows, I will use the term “thought” only with this loose meaning. It should not be understood as the technical notion of “thought” of Frege’s mature thinking, i.e., as the *sense* (*Sinn*) of a complete sentence, which is already a substantive thesis about its nature.

reasoning is sometimes unwarranted or unjustified: we are justified in moving from one thought to the other if and only if this move is in accordance with the most general laws of thought. In this sense, then, logic is also a *normative* discipline, for it sets the standards for all rational thinking: it prescribes how we *ought* to think if we want to think rationally\(^\text{17}\).

This, however, does not explain the real value of logic, for the same laws apply for true thoughts as well as for false thoughts, and nobody is really interested in false thoughts. Frege believed that logic is valuable because, in uncovering the laws of thought, it also uncovers the laws of *truth preservation*: if a given thought is true, then all thoughts that are appropriately connected with it are also *guaranteed* to be true. And arriving at truths is the goal of all sciences. This is why logic is so important: it lays down the rules that must be followed by all sciences if they want to be guaranteed to arrive at truths.

In this way, then, logic sets up the ultimate grounds of justification for any science, and hence for all knowledge. If a given chain of deductive reasoning follows precisely the natural ordering of truths, nothing more can be asked of it in terms of justification – we are fully justified in believing its conclusion. This is exactly what a proof is: a chain of deductive reasoning that lays down, step by step and in conformity with the laws of truth, the systematic transition from the thoughts that constitute the premises to the thought that constitutes the conclusion. In other terms, a proof is something that “reveals the *logical self-evidence* of its conclusion” (Dickie, 2008, p. 275, emphasis mine) by showing its relation to the premises in *logically self-evident* steps, which is the same to say that they follow the natural order of truths\(^\text{18}\).

It is also important to stress that Frege believed that thoughts are *structured*, i.e., that they are build out of simpler parts that refer\(^\text{19}\) to or are about objects and properties, and they obey a certain mode of combination. These parts, of course, can appear in any number of thoughts, combined in different ways. Different thoughts, then, may have parts in common. Moreover, every thought component (whatever they are) has what I will call, following Dickie (2008), *inferential properties*, i.e., properties that affect how the thoughts they are components of relate to other thoughts. In different terms, every

\(^{17}\) See McFarlane (2002) for a discussion of Frege’s conception of logic as a descriptive and normative discipline. See also Frege (1997), pp. 227-250.

\(^{18}\) This point is very well explained by Dickie (2008). See also Frege (1960b), p. 102 for his discussion of proofs and of logical self-evidence.

\(^{19}\) I am using the term “refer” here in a very loose sense, meaning only that these thought parts are about things in the world.
relevant thought component, whether it refers to objects or properties, has certain properties that explain why the thoughts that contain them lie where they lie in the natural order of thoughts. So, one thought follows naturally from the other because their components have suitably related inferential properties. The inferential properties of thought components, then, determine the inferential properties of complete thoughts, i.e., which inferences are authorized or warranted, and are what explains the transmission of truth from the premises to the conclusion in a deductive reasoning. Hence, to study thoughts and the laws that govern them, the logician must study their structure and their parts, and explain how they have the inferential properties they have. Only in doing so she can account for how some thoughts may be legitimately inferred from others whereas some may not.

But how, after all, can she study thoughts? They are immaterial and abstract things which cannot be put under a microscope for inspection. Frege expressed this concern in a footnote of *Der Gedanke*:

I am not in the happy position here of a mineralogist who shows his hearers a mountain crystal. I cannot put a thought in the hands of my readers with the request that they should minutely examine it from all sides. I have to content myself with presenting the reader with a thought, in itself immaterial, dressed in sensible linguistic form. (…) So a battle with language takes place and I am compelled to occupy myself with language although is not my proper concern here. (Frege, 1956, p. 298, footnote 1)

Now we can understand why Frege was interested in language. His primary interest was in thoughts and their laws, but thoughts are not immediately accessible. The only way to access them and inspect their properties is through their linguistic embodiment. But how can Frege be so sure that language is so safe a guide to thought? As said earlier, thoughts are structured entities that are expressed through language. If thoughts are indeed structured and can be expressed in language, then language must also be structured, at least to a certain degree. As Frege puts it,

even a thought grasped by a human being for the very first time can be put into a form of words which will be understood by someone to whom the thought is entirely new. This would be impossible, were we not able to distinguish parts in the thought corresponding to the parts of a sentence, so that the structure of the sentence serves as an image of the structure of the thought. (Frege, 1963, p. 1)

As we can see, Frege’s interest in language is only derivative. He is concerned with language only as a means to study thought and its laws. If the structure of language
mirrors the structure of thought – at least to some degree – then the logician can investigate language to achieve his goal. However, natural language is too messy a medium for the expression of thought. Of course, its function is to express and communicate thoughts, but it is so filled with impurities and complications that figuring out which thought is being expressed in a given occasion requires a great level of skill and ingenuity. Natural language, then, is “an imperfect vehicle for thought” (Burge, 2005, p. 246), as it does not reveal in a precise way the relevant structure of thoughts and their relations. In fact, Frege thinks that logic is necessary precisely because natural language is ill suited for expressing thoughts accurately. If the logician really wants to study thought in the most fruitful manner, she must remove all the impurities from natural language, in the same way that the scientist isolates her experiment from all outside interference. The best way to do this, Frege held, is to create a new language that is “ideally suited to the expression of thought” (Burge, 2005, p. 245), one that captures only its relevant structure and is able to precisely represent its components and their inferential properties. Moreover, such a language would be extremely useful for other scientific disciplines as well, because it would be much better suited for expressing their discoveries and their justification than natural language is. And developing this ideal language is exactly what Frege did in his *Begriffsschrift*.

Before talking about this ideal language, however, let us take stock. Frege believed that thoughts are related to each other in such a way that some thoughts follow naturally from others. They form a natural order of thoughts, and the job of the logician is to describe the laws of this natural ordering. Uncovering these laws is important because it prescribes how we ought to construct proofs, which are the ultimate justification for all knowledge. Something is a proof if and only if there are no gaps or intermediate implicit thoughts in its chain of reasoning, i.e., if it represents correctly the natural order of the thoughts in question. Thoughts are structured entities which are composed of more elementary parts that are about things in the world, and these parts have properties that determine where the thoughts that contain them lie in the natural order of thoughts. So, to study thought appropriately, the logician has to study their parts and how they relate to each other. Since the structure of language mirrors the structure of thought, then the best way to study thoughts is to devise a perfect language that captures precisely the structure of the thoughts it expresses.
The language of the BS has many interesting aspects, but only one (which I think is often overlooked) will prove crucial for understanding the opening passage of SuB. This aspect has to do with the relation between syntactical form and logical form in this language. As I mentioned, the purpose of the BS was to devise a language that was able to express thoughts and their logical structure clearly and without any ambiguity. Hence, contrary to natural language sentences, no question of which thought is being expressed would arise for sentences in the perfect language, and its logical structure would be immediately apparent. Since what matters for logic are the inferential properties of thoughts and thought components, this language is better able to represent their proper distinctions. The best way to guarantee that this occurs is to stipulate that, in this language, to every syntactical difference must correspond a logical difference. In other terms, in the language of the BS, syntactic representations must always reflect a difference in inferential properties of what is being represented. Hence, in the perfect language, syntactical form and logical form always coincide.

Frege makes the point above about syntactical form because, as he says in SuB, “nobody can be forbidden to use any arbitrarily producible event or object as a sign for something” (Frege, 1960a, p. 57). This means that someone can arbitrarily choose to represent the same thought or thought component with as many different syntactic representations as she likes. For example, I may very well arbitrarily choose to represent the same thought component with the shapes a, b, c and d. These shapes, however, do not represent a logically relevant difference according to Frege; the shapes a, b, c and d would be mere notational variants of the same logically relevant content, in the same way as the words “color” and “colour” are notational variants of the same concept. The inferential properties of what they represent are identical. As it seems clear, such distinction serves no purpose in the BS. If we want a language that is able to display without ambiguity the logical structure of thoughts, having extra syntactic representations to express the same thought or thought component would only complicate things. In fact, excess of syntactic representations is one of the things that make natural language ill-suited for the accurate expression of thought. For this reason, Frege posits a normative constraint on syntactic representations for the ideal language: different orthographic shapes (such as a and b) must represent different logically relevant contents, so that a syntactic difference always reflects
a logical difference in the thought components they express. Hence, syntax is at the service of logic.

This is precisely what we must have in mind when we read the first paragraph of *SuB*. When Frege talks about statements of the form $a=a$ and $a=b$, he is not assuming anything about the syntactical or logical form of *actual* proper names and sentences of natural language, and claiming that sentences of these forms have different cognitive profiles. If this was the case, Frege would of course need to provide a criterion for name individuation to determine when a given sentence of natural language is of the form $a=a$ or $a=b$, as Glezakos points out. Rather, he has in mind sentences of these forms in *his ideal language*, where thoughts and their logically relevant structure are clearly and unambiguously expressed by the syntactical representations of that language.

So, when he says that sentences of the form $a=a$ and $a=b$ have different cognitive values, he really means that *thoughts* of the form $a=a$ and $a=b$ have different cognitive values: the first represents a trivial or tautological self-identity, whereas the second represents an informative one. And that there *are* such identity thoughts seems clear enough: I can certainly think twice about the same object and know beyond doubt that I did so, and at some point of my cognitive life I certainly was informed by coming to know that what I regarded as distinct objects were in fact the same object. The mere existence of this phenomenon is enough to generate the puzzle for Frege. It does not matter at all which actual, existing sentences of natural language express thoughts of these forms and under which circumstances, or even if any of them do; we can very well *stipulate* in an ideal language which sentences and symbols represent precisely the identity thoughts in question. Thus, Frege does not need to answer if the actual sentence “Paderewski is Paderewski” is of the form $a=a$ or $a=b$ to be able pose his puzzle, or any sentence of natural language for that matter; in his ideal language this question simply does not arise, for the syntactical form of a sentence always reflects the logical form of the thought expressed by it.

If we interpret the schemas $a=a$ and $a=b$ in this way, then what Frege had in mind with the puzzle in *SuB* becomes much clearer. It goes like this: why are there significant *logical* differences between two thoughts that require identical conditions to

---

obtain in order to be true? Why two thoughts that are about the same objects and the same properties can nevertheless differ in their inferential properties? In other words, why are we justified in deducing set S of inferences from one thought while at the same time not justified in deducing set S from the other, if they are extensionally equivalent? Why do they occupy different places in the natural order of thoughts? In its core, then, Frege’s puzzle has little to do with language. It does not arise by first considering natural language sentences and then asking why these sentences differ in their cognitive profile; the puzzle arises for Frege because some thoughts with coreferring parts can have distinct inferential properties, and this fact about thought is completely independent of there being actual natural language sentences that express it. Language comes into play only because there is no way to study thought and its laws other than by clothing it in linguistic form. In order to talk about the puzzle, we must put it down in symbols, but we do not need to borrow natural language sentences to do it. We can now see that Glezakos’ criticism, which reads Frege as claiming something about the syntactical and logical form of natural language sentences, is not a threat to the puzzle posed this way. It simply misses its target.

To sum up. We saw that the puzzle does not arise only for identities, that it has nothing to do with analyticity or a prioricity, and that we should not interpret the schemas \( a=a \) and \( a=b \) in the opening passage of SuB as representing natural language sentences but as representing forms of thought. If we have this in mind, and if we understand Frege’s general views on the connection between thought, logic and justification, then we can appreciate what exactly Frege is puzzled about. In any case, it seems clear that natural language was not his concern, and so he made no assumptions about actual sentence forms.

What Frege needs to assume to pose the puzzle in SuB, of course, is that coreferential thoughts may differ in a logically significant manner, and that “cognitive value” (Erkenntniswerte) is a logical property. But this is precisely what he assumes. Put another way, what is implicit in his formulation of the puzzle is that extensionally equivalent thoughts and thought components can have significant cognitive differences – what he calls cognitive value – and that these cognitive differences must be accounted for by logic. And this is a very reasonable assumption to make, given his views on thought and on what logic is about. Consider, for example, a perfectly rational subject who has two thoughts that can be expressed in an ideal language as \( Fa \) and \( Gb \), where \( F \) and \( G \) represent

---

21 See Taschek (1992) for a very nice discussion of this point.
thought components that are about properties, and \( a \) and \( b \) represent thought components that are about objects. This subject does not know that \( a \) and \( b \) corefer, so there is a cognitive difference between thoughts \( Fa \) and \( Gb \). Even if she is ideally rational, she would not be able to infer that \((F \land G)a\) or \((F \land G)b\) without discovering that \( a=b \), even though these thoughts are all true. Of course, nothing prevents the subject from moving from the thought that \( Fa \) or \( Gb \) to the other conjunctive thoughts for whatever psychological reasons; the point is that she would not be able to do so justifiably and rationally. In other terms, since the subject is ideally rational and cannot tell at the outset that \( a \) and \( b \) are about the same object, it must be the case that \( a \) and \( b \) have different inferential properties, so moving from one thought to the other would not follow the natural order of thoughts, and hence the move would not be justified. Moreover, as Frege says in \( FuB \), a rational subject may regard one thought as true while the other false, due to no fault of her own. It seems quite natural to think that this sort of cognitive difference arises because the relevant thoughts have different logical properties, and this is precisely what Frege believed.

The same example can be put in terms of a simple proof (with the proper rules of inference):

\[
\begin{align*}
\text{i. } & Fa \\
\text{ii. } & Gb \\
\text{iii. } & a=b \\
\text{iv. } & Fb \\
\text{v. } & Ga \\
\text{vi. } & Fb \land Ga
\end{align*}
\]

If we remove step (iii), we would no longer be justified in believing the thoughts expressed in (iv), (v) and (vi), for removing it would transform a chain of reasoning that is built out only of logically self-evident steps into one that is not. Hence, this chain of reasoning would no longer be a proof: there is a logical gap between steps (ii) and (iv) that prevent it from being self-evident, despite the fact that all thoughts contained in that purported proof are true. If this is the case, then it seems clear that the thought components represented by \( a \) and \( b \) have distinct logical properties even though they are about the same object. The proof would likewise be flawed if, instead of using the sentence

\[22\text{ This point is made by Dickie (2008).}\]
$a=b$ in step (iii), we used the sentence $a=a$. Sentence $a=a$ tells us nothing we do not already know. The thought it expresses is trivial, and thus is not connected to the other thoughts contained in the purported proof in a suitable way as to render its conclusion self-evident. In sum, then, the above examples seem to show that the cognitive difference between $a=a$ and $a=b$, and between $a$ and $b$, are logical differences.

Now we can also understand what Frege really means by the mysterious property of extending our knowledge (or informativeness) that he attributed only to $a=b$ sentences as opposed to $a=a$ sentences. Sentences of the form $a=b$, but not sentences of the form $a=a$, can render an otherwise ‘gappy’ and not self-evident chain of reasoning into one that is self-evident, thus justifying us in believing the conclusion. By coming to learn that $a=b$, then, a rational subject can draw inferences that she could not (or was not warranted) before. She can come to see new connections and relations between thoughts that she already believed, and thus might correct (metaphysically) inconsistent beliefs in her belief system that could not be corrected before. Suppose, for example, that our subject believes (correctly) that $Fa$ and $Gb$, and believes (incorrectly) that $\neg Ga$ and $\neg Fb$, and she does not know that $a$ and $b$ are coreferential. She is not able to tell, by logic alone, that she has inconsistent beliefs. By coming to learn that $a=b$, however, she might immediately perceive this inconsistency, even though she already believed that $a=a$ and $b=b$. Sentences of the form $a=b$, then, may cause a reorganization of our beliefs and our knowledge that sentences of the form $a=a$ cannot, precisely by revealing connections that were not knowable in advance. And the revealing of new connections and relations between thoughts is precisely what Frege means by *extension of knowledge*\(^\text{23}\), even if nothing essentially new emerges in this process (Frege, 1960b, p. 100). This is why even analytic and *a priori* sentences can be informative if they are of the form $a=b$: they may reveal connections between thoughts expressed by other analytic sentences that were otherwise not self-evident, even if they do not contain any essentially new (synthetic) information in themselves\(^\text{24}\).

It seems very plausible, then, to assume that the sort of cognitive differences that Frege detected between coreferential thought components or between trivial and


\(^{24}\) See Frege, (1960b), pp. 100-1, where he discusses the capacity of analytic sentences (definitions in arithmetic, to me more precise) for extending our knowledge.
informative self-identities are logical differences. If the activity of rational thinking is the activity of taking possession of thoughts and trying to follow their natural order, as Frege believed, and if an ideally rational subject is simply not able to draw a certain inference from a given set of coreferential thoughts, then her inability to do so must be due to a logical difference in those thoughts and their components. Their inferential properties are simply not suitably related as to reveal the transition from the premises to the conclusion. Even if she does move from the premises to the conclusion with a ‘gappy’ reasoning, she would not be justified in doing so, because this move does not follow the natural order of thoughts. Frege’s view of rationality, then, is inextricably connected to logic and justification, and Frege’s puzzle lies exactly at the intersection of these issues.

This is why solving this puzzle is so important for Frege. He knows (or reasonably assumes) that there are logical differences between coreferential thoughts, but we do not know why these differences arise. In other terms, he does not know why extensionally equivalent thoughts can have distinct cognitive values. Without a proper account of why this is the case, the logician cannot offer a complete account of thought, its laws and their relation to rationality and justification. It is not enough to describe the laws that govern the transitions from one thought to the other. She must also give an account of what endows thoughts with the inferential properties they have, i.e., what makes them have the logical forms they have. Put another way, she must explain what our reasoning is sensitive to, and hence explain why coreferential thoughts may differ in cognitive value. Only in doing so she will be able to offer a complete account of what constitutes a proper proof and hence of what is a proper justification for scientific knowledge.

When the puzzle first appeared, in BS, Frege had inconsistent views on this matter. On the one hand, Frege apparently believed that what was responsible for endowing thoughts (and their components) with the inferential properties they have were only the relevant circumstances (Umstände) whose obtaining or not obtaining made them true of false (Taschek, 1992, pp 769-770). In other words, he believed that what determined the relevant inferential properties of thoughts were simply their referential truth conditions25. Moreover, he apparently held that what was relevant for determining the inferential properties of the part of a thought that is about an object was the object itself: if

25 Cf. Taschek (1992), pp. 769-770 and Mendelsohn (2005), p. 43. See also Frege (1960a), pp. 2, 5, 6, 7, for examples of his use of “circumstance” (Umstand).
two thought components are coreferential, then they have the *same* inferential properties, i.e., they are logically identical. Hence, two thoughts with the same referential truth conditions would be identical for the purposes of the *BS*: they would occupy the same place in the natural order of thoughts and could be indifferently interchanged in a proof. In this aspect, Frege’s position is very Millian in spirit, for the bearers of logical properties in the *BS* seem very much like singular propositions, which have objects themselves as ingredients. I will call this “the referential view” of inferential properties.

On the other hand, Frege already had the intuition that our reasoning is not sensitive only to the referential truth conditions of our thoughts, but also to the *modes of determination* (*Bestimmungweisen*)\(^{26}\) of these truth conditions; more specifically, to the modes of determination of the relevant objects. Frege explicitly claims that “we *must* supply, corresponding to the two ways of determination, *two different names* for the thing thus determined” (Frege, 1960a, p. 12-3, emphasis mine). As I mentioned earlier, one of the important aspects of the ideal language is that, to avoid ambiguity, there must be a one to one correspondence between syntactic representations and logically relevant contents. When Frege says that we *must* use different names when an object is determined in different ways, it is seems clear that he is under the impression that modes of determination are responsible for endowing thoughts and their parts with their inferential properties, and hence we should represent this fact in the ideal language by supplying distinct names for the same object. As he says, giving different names to the same object is “not always just a trivial matter of formulation; if they go along with different ways of determining the content, they are essential to the nature of the case” (Frege, 1960a, p. 12). In fact, he even says a bit earlier that we are not *justified* in using the same name for an object determined in different ways; only a proof that reveals their equivalence could authorize us to do so. These claims would make no sense if he did not already had at least the intuition that what is relevant for thought and its logical properties were not the objects it is about, but their modes of determination. I will call this “the mode of determination view” of inferential properties.

This is why an identity sign cannot be thrown away in a perfect language. On the referential view, if two terms \(a\) and \(b\) are coreferential, then they have the same

\(^{26}\)This notion is equivalent to his later notion of *mode of presentation* (*Art des Gegebenenseins*), but I opted to stick with “mode of determination”.
inferential properties, and hence could be interchanged in a proof or in any chain of reasoning without compromising the logical self-evidence of its steps. They have, as Frege would later put it, the same cognitive value. Thus, there would be no logical difference whatsoever between sentences of the form \( a = a \) and \( a = b \), and the identity sign would be completely dispensable. But Frege is eager to show that this is mistaken. He says that, as soon as the terms \( a \) and \( b \) are joined by an identity sign, they “appear in propria persona” (Frege, 1960a, p. 10), and thus represent themselves in the statement. The sentence \( a = b \), then, expresses something like “the name ‘a’ has the same content as the name ‘b’”. And this is not something irrelevant, for, as we saw, names in an ideal language go along with different ways of determining the content, and hence identity statements are able to express significant knowledge about objects.

Clearly, the referential view and the mode of determination view are incompatible. Either our reasoning is sensitive to the referential truth conditions of the thoughts we grasp – and is therefore sensitive only to their referents – or it is sensitive to their mode of determination. It cannot be both. If \( a \) and \( b \) are coreferential, they should have the same inferential properties according to the referential view; it should not matter if they are determined in different ways. We could use either one in a proof and the proof would still be logically self-evident. According to the mode of determination view, however, it does not matter if \( a \) and \( b \) corefer; if they are determined differently, then their inferential properties also differ. Hence, we cannot substitute one term for the other in a proof without rendering it not self-evident. One of these views must be given up.

Frege probably failed to see the incompatibility in his BS views because he thought that modes of determination were important only in the context of identity statements, and he managed to deal with the problem with his metalinguistic solution. He did not see that substituting one term for the other could change the inferential properties of any sentence, not only of identity statements, and could thus invalidate otherwise self-evident proofs. By the time of SuB, however, Frege came to realize that the puzzle is general, and so the referential view must be abandoned. He realized that what it is really relevant for endowing thoughts with their inferential properties are the modes of determination of the objects we think about, not the objects themselves. In other words, he recognized that objects are not sufficient to determine the logical properties of thoughts that are about them. So, in order to construct adequate, logically self-evident proofs which
follow the natural order of thoughts, we must pay attention to the mode of determination of objects, and not only to the objects themselves. We must distinguish between the *Bedeutung*, which is the extension of thoughts and their parts, and the *Sinn*, which is the stuff that thoughts themselves are made of, and wherein the mode of determination is contained. It is the *Sinn* that explains why two extensionally equivalent thoughts can nevertheless differ in their inferential properties. It is the *Sinn* that determines where a given thought lies in the natural order of thoughts, not its *Bedeutung*, and therefore it is the *Sinn* that must be represented in a proper proof.

To summarize: Frege detected a cognitive difference between coreferential (or extensionally equivalent) thoughts, and believed that this difference can only be a logical difference. The question he sets himself to answer is why this logical difference arises. In the *BS*, he had conflicting intuitions on this regard; by the time of *SuB*, however, he was clearly aware that objects themselves cannot determine the inferential properties of thoughts that are about them. In other terms, the *Bedeutung* was not sufficient for logic and for explaining cognitive value. He needed the *Sinn*, and modes of determination, to explain why this logical difference between coreferential thoughts arises.

### 1.3 Are the puzzles different after all?

As we can see from the foregoing discussion, Frege’s version of the puzzle seems considerably different than the problem that traditionally goes under the name of “Frege’s Puzzle”, which was described in section 1. If this interpretation of Frege is correct, he saw the puzzle as essentially a *logical* problem about thoughts and their relation to proof theory and rationality, with nothing to do with actual natural language sentences (or utterances) and their impacts in the cognitive life of speakers. But are the two versions really so different?

It seems to me that they are not. Both versions deal with the same general phenomena – i.e., the phenomenon of trivial and informative self-identities, and the phenomenon of differing cognitive values of extensionally equivalent thoughts (understood in a very loose manner) – but from different perspectives. Frege was not really interested in natural language; more precisely, he was interested in language only insofar as language can help understanding thought and its logical properties. Natural language, for him, then, occupied an extremely marginal place in his philosophical project. He never really cared
about working out the details of how the problematic thoughts he was concerned with were actually expressed in natural language. He also assumed that the phenomena of cognitive value were fundamentally logical phenomena, i.e., that they should be accounted for by the science of logic. Because of this assumption, he placed the puzzle at the center of his broader philosophical project of devising a perfect language in which we could adequately carry out proofs, and hence provide the ultimate grounds of justification for scientific knowledge. Hence, given Frege’s background assumptions, the general puzzle assumed a much more specific form and a much more specific role in Frege’s philosophical endeavor.

The puzzle as presented in section 1 involves the same phenomena of cognitive value that Frege recognized, but it takes natural language as its starting point. To pose that version of the puzzle, the one which is traditionally labeled “Frege’s Puzzle”, we do not need to assume any substantial thesis about logic, rationality and justification of scientific knowledge as Frege did. We need only to (a) recognize that the phenomena of triviality, informativeness and cognitive value of thought (understood broadly) do exist, and (b) ask ourselves when and how these phenomena can be manifested in natural language. In other words, we just need to acknowledge the existence of such phenomena and acknowledge that natural language can, at least sometimes, exhibit them27. We can then set ourselves to the task of explaining why this is the case, independently of any prior assumptions about logic, semantics, or whatever. For it seems quite evident that natural language can, in fact, manifest these phenomena: at some point of our lives we certainly apprehended an informative identity statement, and we can easily think of situations where natural language sentences can be used to express trivial self-identities. To engage with this puzzle is to explain under what circumstances this is possible, and why. And a great variety of explanations have been proposed.

In sum, then, both the general puzzle from section 1 and Frege’s version of the puzzle deal essentially with the same general phenomena about thought and cognitive value. The difference lies in their background assumptions and in their starting point, and what they think that cognitive value is. This finally brings us to Taschek’s criticism of direct reference theorists and their approach to the puzzle.

27 See Boccardi (2014) for the claim that Frege’s Puzzle can be likened to the problem of de jure (or explicit) coreference vs. de facto coreference.
Taschek’s criticism of direct reference has to do with the issues about inferential properties of thoughts and referential truth conditions I mentioned earlier. He also interprets Frege more or less along the lines of what I have presented here, and claims that the real challenge of Frege’s puzzle is precisely to “account for the way in which the logical properties of the contents [i.e., thoughts] of our sentences can outstrip their referential truth conditions” (Taschek, 1992, p. 788). And Frege met this challenge by rejecting the referential view and accepting the mode of determination view of inferential properties. According to Taschek, even if Frege’s theory of Sinn and Bedeutung is ultimately flawed, it is certainly in keeping with our most intuitive view on the role that logic has in our cognitive and linguistic practices and in the way we intuitively evaluate beliefs in terms of their implications, their consistency, etc. As I said earlier, it seems very plausible to assume that, if an ideally rational subject cannot draw some inference or does not realize that she has inconsistent beliefs due to no fault of her own, then it must be the case that her beliefs differ in a logically significant manner: her beliefs may be metaphysically inconsistent, but not logically inconsistent. We cannot explain this fact by resorting only to their referential truth conditions, and Frege noticed this. In sum, then, Taschek believes that Frege is right in associating logic and rationality the way he does.

Direct reference theorists, on the other hand, reject this association without providing sufficient arguments, according to Taschek. They deny, even if implicitly, that rationality implies logical consistency. For them, what is relevant for determining the logical properties of the contents of our beliefs is just their referential truth conditions. In other words, direct reference theorists hold a referentialist view of inferential properties, which is precisely the view that Frege saw the need to eschew when he faced the puzzle in SuB. On direct reference, then, a subject that is ideally rational can nevertheless have logically inconsistent beliefs, as is illustrated by poor Pierre: he does believe a contradiction when he believes what is expressed by the sentences “London is not pretty” and “Londres est jolie”. Yet, he is not irrational. If it is possible for a rational subject to indeed believe a contradiction, then direct reference urgently owes us an adequate account of the role that logic plays vis-à-vis rationality and vis-à-vis our ordinary evaluations of consistency and implication in a subject’s system of beliefs. As Taschek puts it,
In other words, Taschek thinks that it is not enough to brandish singular propositions against Frege; they may very well refute traditional Fregean semantics, but to really engage with his puzzle, direct reference theorists must occupy themselves with deeper issues about rationality and logical consequence. The solutions they offered to the puzzle presuppose substantive theses about these questions, yet they are far from being clearly articulated and well defended. Direct reference theorists, then, may have dealt with one aspect of Frege’s puzzle, but failed to tackle what really needs to be tackled. But is this criticism of direct reference appropriate?

Yes and no. It is appropriate in the sense that major direct reference theorists, such as Kaplan and Perry, as far as I am aware, never interpreted Frege’s puzzle precisely in this manner, and because of this they never systematically addressed these issues directly in connection to the puzzle. They faced it mainly on semantic grounds. However, there is a vast literature on the subject of content externalism, rationality and logical consequence. If the moral of Frege’s puzzle is really about rationality, logic and its normative constraints in our actual reasoning, then it seems clear that the problem is being discussed appropriately, even if it is not directly associated to the puzzle itself. However, I believe that, even if direct reference theorists did not engage with Frege’s version of the puzzle, they have clearly engaged with the general puzzle I presented in section 1, which as we saw are not so different after all. Their focus is natural language, and their broad philosophical projects are very different than Frege’s. For this reason, they cannot be blamed if they did not occupy themselves with precisely the same issues which Frege believed were so pressing. Of course, a complete account of the puzzle and of all its aspects and implications for thought and language must at some point tell us where logic fits in, and what role it plays. But I believe that logic is by no means the only relevant aspect of the puzzle, and is not the most important. It certainly was for Frege given his background assumptions, but it does not need to be for us who are mainly interested in

---

28 Taschek’s criticism is, in some aspects, very similar to Boghossian’s (1992) criticism of content externalism.
natural language. In a word, we do not need to assume that “cognitive value” is a logical notion, as Frege did.

1.4 Concluding remarks

The puzzle for Frege and the puzzle that has been attributed to him are somewhat different, but they are based on the same interesting phenomena of cognitive value and thought (understood in a broad sense). In any case, even if Frege’s version of the puzzle has some distinctive particularities, Frege has been widely interpreted as setting an adequacy criterion for semantic theories, namely, that semantic theories must account for the phenomena of cognitive value. Even philosophers who abandoned Fregean semantics, such as Perry and Kaplan, believed that this adequacy criterion is correct. They believed that cognitive value is an aspect of meaning, so they tried to account for it on semantic grounds. In what follows, I will argue that their attempts to do so fail, and that this gives us very good reasons to suspect that cognitive value is not an aspect of meaning. From now on, all mentions of the puzzle will refer to the puzzle as arising for natural language, the one that direct reference clearly engaged with, not to Frege’s more particular version.
2. Indexicals.

As I mentioned in the first chapter, Frege’s Puzzle does not arise only for sentences involving proper names. It also appears in sentences involving indexicals. As we will see, indexicals are the most appealing starting point for the direct reference theorist who is sympathetic to the Fregean approach to semantics. In this chapter, I will explain how Frege’s Puzzle arises for indexicals and why the solution via character and via reflexive content do not work, even in more sophisticated forms. To do that, I will first explain what indexicals are and present the Kaplanian semantics, which is the orthodox theory of indexicals. In order to deal with some naïve objections that can be posed to Kaplan’s solution to the puzzle, I will first present his theory in a somewhat simplified form (as Wettstein (1986) apparently understood it), and then in its strongest version. After dealing with the solution via character, I will explain what reflexive content is and argue that it also suffers from the same problems as a modified version of the character.

2.1 Pure indexicals and demonstratives

Indexicals are linguistic expressions that have different semantic values depending on the context of their utterance. The clearest cases of indexical words are words like “I”, “you”, “me”, “here”, “now”, “this”, “that”: depending on the occasion of their use, they refer or designate different things. For example, I might utter the sentence “it is raining here” in Campinas and Kim Jong-un might utter “it is raining here” in Pyongyang. We say different things (I may utter a falsity while Kim Jong-un utters a truth) by using the same sentence. The explanation for this phenomenon is that the referent of the word “here” depends on the context of its use. When I utter “here”, I am referring to Campinas, and when Kim Jong-un utters “here”, he refers to Pyongyang. The same sentence is used to express different propositions. The same phenomenon occurs with all other indexical words. We call this feature context sensitivity. Many other words exhibit some kind of indexicality, like “foreigner” or “neighbor”\(^{29}\), but they do not concern me here. I focus only on the most emblematic examples of indexicals, which were the most extensively discussed in the context of Frege’s Puzzle.

The first important thing to notice is that indexicals are not a monolithic semantic category. Kaplan (1989a) distinguished between pure indexicals and demonstratives\(^{30}\). Pure indexicals are words like “I”, “here”, “now”, “today” and “yesterday”. They are indexicals that do not require an accompanying gesture or pointing in order to secure a referent in the context. This means that when I utter an indexical of this sort I do not need to perform any kind of demonstration in order to determine a referent in the context. When I say “today is a horrible day”, I am not required to point nor do any kind of gesture to indicate to which day I am referring. It is clear that the mere utterance of the word “today” in the context is enough to determine the day in question. The same thing happens with “I”. I do not need to point at myself in order to convey which object I am referring to. Simply by hearing (or reading) me saying “I” is sufficient for the hearer to know that I intend to refer to myself. Of course, an utterance of “I” may be accompanied by some gesture, like when I angrily say “I was the one who did it!” while repeatedly pointing at my chest. However, this pointing is merely a resource for emphasis, not a necessary device for determining the referent. What I said would be equally understood even if I had my hands tied to a chair. Any demonstration accompanying a pure indexical, therefore, is “either for emphasis or irrelevant” (Kaplan, 1989a, p. 491).

On the other hand, demonstratives are not capable of securing a referent by themselves. They are words like “this”, “that”, “he”, “she” and “there”, that require some sort of demonstration in order to determine a referent in a context\(^{31}\). Imagine a situation in which I am walking down the street with a friend and we spot three little boys coming toward us. I recognize one of them as my most annoying neighbor. I then turn to my friend and say “he is a little monster” without any pointing or gesture whatsoever. It seems obvious that my friend would be at a loss to know to which boy I am referring. The word “he” by itself does not do any discriminating job in this context: there are three boys to which I may be referring, and none of them is particularly salient. In order to specify for the audience which boy I have in mind, I have to point at him (or tilt my head to be more discrete) in addition to uttering the indexical “he”. It is the same with all other demonstratives. An utterance of “that” or “this” without any demonstration will be, as Kaplan puts it, incomplete, for the words by themselves are incapable of securing a

\(^{30}\)Most philosophers accept such distinction. The only apparent exception that I am aware of is Salmon (2002). He believes that pure indexicals and demonstratives are both complete in the relevant sense, so no significant distinction is warranted, at least not at the level of meaning.

\(^{31}\)I am not considering cases of anaphora or bound variables, just the demonstrative uses of these words.
referent, i.e., distinguishing one object from others in the context. In summary, demonstratives are the type of indexical that requires an accompanying demonstration in order to refer successfully, while pure indexicals can do this by themselves.

It is important to stress that the notion of “accompanying demonstration” is not restricted to actions performed by the speaker. Kaplan (1989a) believes that any sort of feature of the context that makes an object sufficiently salient is enough to count as a demonstration. Consider the situation described above, but where my annoying neighbor is singing very loudly and drawing a lot of attention. If I say “he is a little monster” without any gesture, my friend would be able to recognize the referent of “he”. There is a feature of the context (the boy’s singing) that makes the referent appropriately salient. I, as a competent speaker, can exploit this fact to refer to him without performing any discriminating gesture. As Kaplan himself says, his notion of a demonstration is “a theoretical concept” (Kaplan, 1989a, p. 490 footnote 9). In any case, we may for now leave open the question of what exactly constitutes a demonstration.

2.2 Character and content

In Kaplan’s theory, indexicals have two types of meaning. One is what he called the character of the indexical, and the other is its content. The distinction between these two types of meaning is fairly intuitive. As we saw, the referent of an indexical shifts from context to context (more precisely, its semantic value). However, it seems evident that despite this shift of semantic value there is something that remains stable in every use of an indexical. In every utterance of “he”, for instance, there is something that keeps constant, despite the fact that this indexical can be used to refer to many different people.

This “something in common” present in all utterances of an indexical expression is what Kaplan called the character of the indexical, the linguistic rule that is attached to the expression. Since linguistic rules of expressions are fixed by the linguistic conventions of language, it is plausible to define character as the linguistic meaning of the indexical. Most importantly, the character is the semantic feature of indexicals that

---

32 Kaplan later changes his mind in Kaplan (1989b). He claims that what completes a demonstrative is a directing intention. More on demonstrations below.

33 We will see, however, that this definition might be mistaken. There seem to be good reasons not to equate character with linguistic meaning. Kaplan himself sometimes seems to hesitate in identifying the two. Cf. Kaplan (1989a), pp. 505, 520-1, 523-4; (1989b), 568, 577-8.
determines the content in a context and it is associated with indexicals as types, not tokens or utterances. In other words, the character is something like an instruction that is attached to the word as a type (somewhat like a dictionary definition) and which “guides” the speakers and hearers to the referent in every situation in which it is uttered. When an indexical is used, it is its character that captures the object being referred to in that occasion. Thus, the character of “he” would be something like “the discriminated male”, and it should be known (even if tacitly) by every competent speaker. In Kaplan’s words:

The character of an expression is set by the linguistic conventions and, in turn, determines the content of the expression in every context. Because character is what is set by linguist conventions, it is natural to think of it as the meaning in the sense of what is known by the competent language user. (Kaplan, 1989a, p. 505)

To make it clearer, we can summarize the properties of character in Kaplanian semantics as follows:

**Character**

1. Determines the content in a context of use
2. Is the linguistic meaning of indexicals
3. Is a property of expressions as types
4. Is what is known by the competent speaker

Content, on the other hand, is what is determined by the character in the context, i.e., its semantic value\(^34\). In this case, the content of an indexical expression is what is referred to in the occasion of its use. When I utter “he is fat”, for instance, the content of the indexical “he” will be the object being demonstrated by me. It is the referent of the indexical that is its semantic value (its contribution to the truth conditions of the sentence), not its character. This, of course, is the fundamental difference between Fregean semantics and direct reference: it is the object itself that constitutes the semantic value of the indexical, not some qualitative content such as the character. Indexicals, therefore,

---

\(^34\) This exposition is obviously simplified. Kaplan says that not only indexicals have character and content, but also all other linguistic expressions (words and sentences alike). The point is that in indexical expressions the character is not identical with its content, while in all other linguistic expressions (non-indexical expressions) the character = content (Cf. Kaplan (1989a), p. 507, 533). In addition, character is compositional, i.e., the character of the parts determines the character of the whole expression. So, for example, the character of the sentence “she starts her new job tomorrow” will be composed by the character of all indexical and non-indexical components, which in turn will provide the content in the context.
express singular propositions (propositions that have objects as their constituents) whenever they are used.

The distinction between the two types of meaning of indexicals seems to capture the idea of shifting content as well as the constant linguistic meaning: the character is given by the linguistic rules and hence remains constant in every use of the indexical, while the content is the object determined by the character in a context. So, we have the following picture:

\[
\text{Character} + \text{context} \rightarrow \text{content}
\]

Having all this in mind, let us see an example of Frege’s Puzzle as arising for sentences involving indexicals and how it supposedly can be accounted for in terms of character.

2.3 Indexicals and Frege’s Puzzle

Consider the following case. I am at a family reunion and I am showing an old photograph album to my aunt. I then point at one ugly and skinny boy at the corner of a picture and say “I am he”\(^{35}\). She gets really surprised, for I am now much different from the kid in question and much better looking. Despite the referent being the same for both “I” and “he” my aunt obviously learned something new after hearing this sentence that she did not know before. The case would be completely different if I said “he is he” pointing at the picture twice, which she would find trivial and uninformative (and rather idiotic). How can her surprise and the informativeness of the sentence be explained by direct reference theorists? In other terms, why the cognitive value of “I am he” is different from the cognitive value of “he is he”?

Clearly this cannot be explained in terms of the content expressed, of what I said. To recapitulate, direct reference theorists, being Millian heirs, hold that the semantic value of an indexical is just its referent. This being so, the proposition expressed in this case is simply the self-identity \(a=a\), for both “he” and “I” refer to the same object: me, Filipe. Singular propositions, as I said previously, are not adequate for explaining cognitive

\(^{35}\) I am ignoring the possible complications of this case. It is not plain obvious that the deferred demonstrative involved in this case (pointing to a picture but purporting to refer to the person depicted) functions exactly in the same way as the demonstrative reference to a flesh and blood person in my vicinities. However, in this case, I think it is harmless to assume that it does.
value. However, there is a new notion available to direct reference theorists that can explain my aunt’s surprise, and it does so in a very elegant way.

Recall the properties (ii) and (iv) of the character mentioned above:

ii) Is the linguistic meaning of indexicals
iv) Is what is known by the competent speaker

Item (iv) is crucial. If a competent speaker must know the character in order to use an indexical successfully (after all, it is the meaning of the expression), then it must play some sort of cognitive role in each use of the indexical. The speaker must have it in mind in one way or another in order to refer to the intended object in the appropriate way or to understand what it is said. If the character already has this epistemic dimension, can it explain cognitive value? Kaplan (1989a) and Perry (1993a) believe that it can.

Let us take a look at what would be the characters of “I” and “he”. The character of “I” could be something like “the person who produced this utterance” or “the present speaker”, whereas the character of “he”, as said above, could be “the discriminated male”. It is clear that the two characters are very different, but both are presented or defined in the form of definite descriptions. In fact, it hardly seems possible to do so otherwise. But what is the relevance of this? Let me explain. Descriptions express conditions that must be met in order for an object to be the description’s satisfier. So, any description will express a condition C, and only if the object satisfies the condition C it will be the object denoted by the description. In other words, descriptions present an object in a certain way, as the object that is so and so, as the satisfier of a certain condition. Descriptions express – to borrow Fregean terminology – modes of presentation of objects, which means that they provide some cognitive perspective over the objects they describe.

Characters, by being some sort of description that must be mastered by any competent speaker, do the same thing; namely, they present the referent under some mode of presentation, under some perspective. Of course, they are linguistic modes of

---

36 Kaplan’s and Perry’s terminology regarding indexicals differ slightly. What Kaplan calls “character”, Perry calls “role”, for example. However, the theoretical role these notions play are the same, so I stick to Kaplan’s more famous terminology.

37 I say that the character “could be” this or that because its exact formulation for each indexical is a matter of debate. It is enough if we try to capture some intuitive definition.
presentation which are fixed by the conventions of language, but are modes of presentation nevertheless. Both characters of “I” and “he” – “the present speaker” and “the discriminated male” – express certain conditions that an object must satisfy in order to be the referent of “I” and “he”: be the present speaker and be the discriminated male. Because of this, these indexicals necessarily present their referents under some mode of presentation, as the present speaker or as the discriminated male. Most importantly, the conditions expressed by characters must be grasped by every competent speaker. When a speaker intends to use the word “he”, for example, she must know that the object she is referring to satisfies the condition of being the discriminated male – and so must the audience.

This is why a sentence like “I am he” can be informative. Each indexical presents the referent in different ways: the indexical “I” presents an object as the present speaker while the indexical “he” presents an object as the discriminated male. Those of course are distinct conditions that could be satisfied by distinct objects. In the context described above, however, the satisfier of these conditions is one and the same: me, Filipe. My aunt, by being a competent speaker who knows how to use and interpret utterances of “I” and “he”, grasps that the two modes of presentation present the same object. In other words, given her knowledge of language, she understands that I am being referred to twice under different (linguistic) modes of presentation, i.e. that I satisfy the two distinct conditions associated with the indexicals. This is the reason why she can become surprised and informed; the referent is the same, but it is being referred to in different ways.

So, in technical terms, the same content can be presented and apprehended under different characters. As we saw, the same content (Filipe) can be apprehended under the character of “I” and under the character of “he” in a context. This makes the sentence “I am he” informative in opposition to the trivial “he is he”, which presents the referent twice in the same way. This applies to all other possible contents that can be expressed by indexical sentences. In the example given above the content expressed was a singular proposition of identity, but it does need to be so. Any kind of singular proposition expressed by indexicals can be grasped under different characters. In short, characters are ways of accessing the contents expressed by indexical sentences. Moreover, since characters are the linguistic meanings of indexicals, they are already given by the rules of language, which any competent speaker should master. Because characters are capable of
acting on the speakers’ cognition, they can explain cognitive value. Different characters, then, can cause different behaviors toward the same content. Let us see some other examples to make this clearer:

**Perry’s supermarket** (Perry, 1993b): Suppose I am shopping at a supermarket when I see in a distant mirror that a man is leaving a trail of sugar wherever he goes. I then think “he is making a mess”. To warn the man, I speed my cart in the aisle, but then I realize that the man in the mirror is myself. I then think “I am making a mess”, so I start to check the things in my cart. I grasp the same content twice, but my behavior is very different because I grasp it under different characters.

**Perry’s bear** (Perry, 1993a): I am in my yard with a friend when a huge bear approaches him. He screams “I am about to be attacked”, rolls up in a ball and stays as still as possible. I, more or less at the same time, scream “you are about to be attacked” and run inside to grab my shotgun. The same content is expressed by both sentences: the singular proposition containing my friend and the property of being attacked. However, he grasps it under the character of “I”, which leads him to a certain behavior, while I grasp it under the character of “you”, which leads me to a different behavior.

**Kaplan’s pants** (Kaplan, 1989a): I see in a window’s reflection that a man has his pants on fire. I laugh saying “his pants are on fire”. After perceiving a disturbing smell, I notice that the man in the reflection is me. I then say “My pants are on fire!”, so I run for help. Same content, different characters, different responses.

In all those cases, the same content – the same singular proposition – Is being expressed or grasped twice. However, this content is being expressed or apprehended under different characters, under different modes of presentation. Those modes of presentation are what affect the cognitive life of speakers, leading to different responses and different behaviors toward the same content. Characters, therefore, individuate cognitive value: *different character, different cognitive value*. Frege’s Puzzle is then solved.
To sum up, contents are not able to account for cognitive value because they are too coarse-grained to individuate cognitive values. This is no news, for we saw earlier that pure Millian semantics seems ill suited in principle for solving Frege’s Puzzle. Characters, on the other hand, explain quite well why certain sentences involving indexicals are trivial while others are informative. Each indexical’s character provides a type of cognitive perspective over its referent, which can lead to different speaker responses. Hence, it is the character that bridges the gap between pure referentialist semantics and epistemology, thus conforming direct reference to the Fregean approach.

It is clear that this discussion has a very obvious Fregean flavor. Characters in Kaplanian semantics perform two of the roles performed by Fregean senses: they are the modes of presentation of the referent and what the competent speaker grasps when he understands an expression. The trick that direct reference theorists did is that they managed to keep these ‘Fregean senses’ out of the propositional content, avoiding all the problems that caused the reaction against Frege in the first place. Characters, therefore, are like Fregean senses, but the cognitive perspective they provide is kept only at the level of linguistic meaning and it is not, to use Wettstein’s phrase, absorbed by the proposition as in Fregean semantics.

Now it is easy to understand why indexicals are the ideal starting point for the semanticist who endorses the Fregean model of semantics. As we saw, an answer to the problem of cognitive value seems readily available through the notion of character. What is important is that this notion was not introduced specifically with Frege’s Puzzle in mind; it was not a somewhat artificial or ad hoc move desperately done by direct reference theorists to explain cognitive value. It feels very natural. Characters were introduced to account for the semantics of indexicals and it seems almost a coincidence that they can also answer the puzzle with such plausibility. In other words, little effort is required to explain cognitive value in the case of indexicals. Once the distinction between character and content was made (with the needs of direct reference in mind), the rest followed more or less naturally.

2.4 Counter examples

Yet, not every direct reference theorist accepted the solution of Frege’s Puzzle via character. Howard Wettstein was one of them. In Wettstein (1986), he argues that Kaplan’s and Perry’s solution only seemingly explain cognitive value. If we consider some
other more problematic examples, he thinks, it will become clear that the explanatory power of character is not as strong as it seems:

**Wettstein’s neo-Fregean attack** (Wettstein, 1986): imagine two utterances of “He is about to be attacked by a neo-Fregean”, but neither the speaker nor the audience know that the same individual is being referred to twice. In fact, it looks like two different individuals are being referred to. It is obvious that the cognitive value of these utterances is very different, for they can lead to very different responses and behaviors.

**Wettstein’s rock star** (Wettstein, 1986): suppose that a rock singer is performing on a stage and he is wearing make-up and clothes in such a way that his left side is completely different from his right side. I then say “he is he” to someone, but in a way that the first “he” is uttered from location A while the second “he” is uttered from location B, dragging the hearer along with me. The hearer then learns that the man seen from the right side is actually the same man seen from the left side.

In both cases, the *same* indexical is used twice to refer to the *same* object. Hence, the same linguistic meaning is operating and the same singular proposition is being expressed. However, it is clear enough that in both examples there is a difference in cognitive value. How can this happen if they involve precisely the same character? In Kaplan’s (simplified) theory, as presented above, cognitive value is individuated by the character; sameness of character, then, should entail sameness of cognitive value. Yet, this is not the case. Recanati (1990) also argues along the same lines, borrowing an example from Perry:

**Recanati’s (and Perry’s) ship** (Recanati, 1990): I say “*this ship* [pointing to part of a ship though a window] is a steamer but *this ship* [pointing to part of a ship through another window] is not a steamer”. Unbeknownst to me, both parts belong to the same ship. Kaplan and Perry would be forced to say that I am irrational, because the same object is being presented under the same linguistic mode of presentation (“*this ship*”) and I nevertheless express (and believe) a contradiction. However, it does not seem plausible to say that I am irrational.
These examples seem to show that it is not possible to explain cognitive value via character, for the linguistic meaning is not enough to individuate cognitive value. Again, sameness of character should entail sameness of cognitive value; yet these examples show that this thesis may be mistaken. In fact, it is easy to think of many cases of “he = he”, “that = that” or “this = this”, where exactly the same linguistic meaning is involved and even so there is a difference in cognitive value.

2.5 Demonstrations to the rescue

One could attempt to rescue Kaplan’s theory by stressing the importance of demonstrations in demonstrative reference. In all counter-examples above the distinction between pure indexicals and demonstratives (section 1.1) seems to be overlooked. Recall that demonstratives (“he”, “this”, “that”) are the type of indexical that can only secure a referent with an associated demonstration. Demonstrations, then, are necessary for the demonstrative to be successfully used. Because of this, we should consider the possibility that the difference in cognitive value in the examples above lies in a difference between associated demonstrations. It could be argued that the demonstration together with the utterance is semantically relevant to determining the referent and that it does so in a cognitively significant manner. If this is true, then it seems that we can preserve the idea that characters can account for cognitive value. Kaplan’s and Perry’s solution would be correct after all.

Kaplan himself explicitly says that demonstrations too have a character: “(…) we can associate with each demonstration a character which represents the ‘meaning’ or manner of presentation of the demonstration” (Kaplan, 1989a, p. 527). The reasoning behind this idea is simple. When a demonstration occurs, a pointing, for instance, it presents the object demonstrated under some perspective, in the sense that when we point at something we necessarily do it from a determinate angle and in a determinate way. It necessarily exploits some appearance or other of the relevant object in that context. In fact, Kaplan claims that demonstrations function very similarly to definite descriptions; so similarly that we can express the character of a demonstration as the description “the individual that has appearance A from here now”, where “an appearance is something like a picture with a little arrow pointing to the relevant subject” (Kaplan, 1989a, p. 526).
If demonstrations too have characters and if no demonstrative is complete without a demonstration, then it seems plausible to conclude that the character of a complete demonstrative (word + demonstration) is given by the combination of demonstrative’s character with the demonstration’s character. To test this idea, let us see how it could account for the counter examples above.

Consider Wettstein’s rock star example. There is a difference in cognitive value even when the same indexical – “he is he” – is used twice. However, the two utterances of “he” are produced from very different perspectives. There is obviously a dramatic change in spatial location between the two utterances. If this is the case, then the accompanying demonstrations that must take place in order to secure a referent exploit very different appearances of the demonstratum. In other words, the manner of presentation of each associated demonstration that accompany the demonstratives is radically different, because one exploits the perspective from location A while the other exploits the perspective from location B. In Kaplanian terms, then, the two associated demonstrations have completely different characters: one is “the individual with the appearance A from here now” and the other “the individual with the appearance B from here now”. Since these descriptions are clearly different, they present the referent in distinct ways. Hence, if we look at the complete character of both demonstratives, we will see that they differ significantly. We would have something like

\[ \text{he} [\alpha] = \text{he} [\beta] \]

where \([\alpha]\) and \([\beta]\) are the characters of the associated demonstrations and “he \([\alpha]\)” and “he \([\beta]\)” are the complete demonstratives. The two sides of the identity are different in character, so character individuates cognitive value after all.

The same argument can be applied to the other examples. They all involve the same demonstrative used twice, but their accompanying demonstrations differ in a significant manner. If this is the case, then the character of each complete demonstrative flanking the identity sign will be different in the end, in accordance with Kaplan’s and Perry’s theory.

In a footnote, Wettstein briefly points out a problem with this move. He argues that “ostending is inessential to the use of demonstratives” (Wettstein, 1986, footnote 17).
and that it is not evident how the above solution would be applied to cases in which there is no demonstrative gesture being performed. However, I think that he did not take into account that Kaplan’s notion of demonstration is not limited to ostension; it is meant to include cases in which no pointing gesture is performed. Of course, the theoretical notion of demonstration offered by Kaplan would need to be developed in order to explain those cases adequately. The point is that, at least in principle, Kaplan’s notion of demonstration is not affected by Wettstein’s objection. However, I think that this solution faces more serious problems.

Kaplan holds that demonstrations should be considered as types\(^{38}\). This thesis may not initially seem very relevant, but I believe it is very important for the argument above. Let us see why.

Remember property (iii) of characters:

iii) Character is a property of expressions as types

If demonstrations were not considered as types, it would not be possible to form a single character compositionally from the demonstration and the demonstrative as the solution requires, or it would seem very implausible to do so. The reason is the following. If demonstrations were considered as tokens, then their characters would be a property of tokens, not types\(^ {39}\). This seems to contradict the idea that characters are general rules of language that have to be mastered, because a character would be formed only circumstantially, at the occasion of the tokening of the demonstration, and would not exist prior to the utterance as Kaplan intended. This view would also be compromised with some weird union of tokens (demonstrations) and types (words). Tokens are physical things, while types are abstract things. How can compositionality be uniformly applied to entities of such different categories? It seems to me that it cannot, at least not in an evident, neat, and non-ad-hoc way. Thus, in order to form the character of the complete demonstrative compositionally, both the character of the associated demonstration and of the demonstrative must be applied to the same kind of thing, i.e., types. Moreover, Kaplan thinks that the same demonstration can be repeated by a different agent and at a different


\(^{39}\) In fact, it seems that character as applied to tokens would be a very different property than the character of expressions as types, for it would involve things like spatial locations and instants of time, and it would be non-repeatable.
place. This would not be possible if demonstrations were considered as tokens, since they would be non-repeatable and non-transferrable. Hence demonstrations must be types, not tokens.

The problem with this idea is the following. If demonstrations are types, then they can be tokened infinitely many times and at different places, as Kaplan thinks is the case. This means that \( n \) number of demonstration tokens can have \textit{exactly the same character}, because they can all be instances of the same demonstration type. How can this explain cases where the same demonstration type is involved, yet the cognitive value of the utterance clearly differs? I will borrow an example from Taschek (1987) to make this clearer. I quote at length:

Suppose, for example, that I am strapped to a chair in a controlled environment and a speaker in the room points to a particular object in the room and says to me:

(2) That is an F.

I believe him. During this time, I correctly believe myself to be in Ann Arbor. Immediately after the demonstration, I am rendered unconscious. For whatever reasons (…) I believe, when I wake up, that a great deal of time has passed, that I am no longer in Ann Arbor, and that I have been moved to Baltimore. I find myself, however, seated in the same place in a room exactly like the one I was in before. Beside me is the same man who was in the previous room. And before me is an object that looks exactly like the one I was shown in Ann Arbor. I think to myself: Ah, they're trying to fool me into thinking that we're still in Ann Arbor! In fact, we still \textit{are} in Ann Arbor; and everything, including the object is exactly as it was before, and only seconds have passed. As soon as I regain consciousness, the man points at the object in exactly the same way as before and says,

(3) That is an F.

But for one reason or another, I do not believe that the object demonstrated in Ann Arbor was moved to Baltimore. This being the case, I do not believe what the man says. I believed what he said when he uttered (2), but, without changing my mind about that, I do not believe what he said when he uttered (3). Thus, (2) and (3) differ for me in cognitive value. (Taschek, 1987, p. 174)

In this example, it seems evident that, if there are such things as demonstration types and tokens, this would be a case where two tokens of the same demonstration type are being performed. The two pointing gestures in (2) and (3) exploit exactly the same perspective, are performed exactly at the same place and demonstrate exactly the same

\textsuperscript{40} Cf. Kaplan (1989a), p. 525-6.
object. In other words, their character is expressed by the description “the individual that has appearance A from here now” and their demonstrata are identical. The only difference is that the two tokens are performed at distinct instants of time; but this is trivial, because two numerically different demonstration tokens cannot be tokened at the exact same place and at the same time. Therefore, there is no distinction whatsoever between the complete characters (demonstrative + demonstration) of the two utterances of “this” in (2) and (3): the demonstrative used is the same and the demonstration type is the same as well. Yet, I regard one sentence as true, while the other as false. Their cognitive values obviously differ to me. How this can be explained if the character is identical in both utterances of “that”?

Those skeptical of Taschek’s counter-example can offer several replies in order to avoid its complications. For instance, one might be tempted to say that the difference in cognitive value arises due to a difference in context: since the contexts in which (2) and (3) are uttered are clearly distinct, their cognitive values are distinct as well, and hence the example poses no threat to Kaplan’s theory. This response, however, is not available to the referentialist who wants to explain cognitive value as an aspect of meaning. The reason is simple: contexts are not in any plausible sense a semantic feature of expressions, as are characters and contents. In other words, contexts are the external things that the meanings of indexicals are sensitive to, which provide the necessary parameters for determining the content of those indexicals; they do not constitute their meanings in any sense in Kaplan’s theory. Meanings and contexts are obviously related, but are totally distinct and independent things. Therefore, under the Fregean conception of semantics, resorting to contexts to explain cognitive value is not an open possibility.

Of course, if one goes Fregean and loads contextual parameters into some level of meaning (as Frege did), one can attempt to explain the above example in terms of some semantic property of expressions: since the contexts of (2) and (3) have different times as parameters, by incorporating these parameters (2) and (3) would then have different meanings, and thus different cognitive values. Clearly, this level of meaning into which contextual parameters could be loaded cannot be the character, because it would then no longer be constant and fixed by the rules of language. However, it could very well be the content. If the contents of (2) and (3) are eternal propositions (i.e., propositions that incorporate times and places), then these propositions would clearly be distinct, for they
would have different instants of time as ingredients. This could perhaps explain the
difference in cognitive value between the utterances that express them. This solution,
however, cannot explain why the same utterances could be trivial: if I was not rendered
unconscious in the example above, and the speaker proceeded exactly as he did, (2) and (3)
would have precisely the same cognitive value despite being uttered in different contexts
and expressing different eternal propositions. If this explanation was correct, a difference
in context (and thus a difference in eternal propositions) should entail a difference in
cognitive value, but this is clearly not the case.

It could be argued that, if I am not rendered unconscious, then the context of
(2) and (3) would be the same, and hence (2) and (3) would have the same cognitive value
after all. My epistemic state seems to affect the context somehow. This argument,
however, is highly implausible. It assumes a notion of context that is not the Kaplanian
one. The context in the relevant sense involves only the parameters necessary to supply the
contents to the expressions in that context. As Kaplan says, “context is a package of
whatever parameters are needed to determine the referent, and thus the content, of the
directly referential expressions of the language” (Kaplan, 1989b, p. 591). As it seems clear,
my mental state and my perceptual tracking of the object have no relevance whatsoever to
providing what is necessary to determine the content in the context: the determined content
is exactly the same whether I become unconscious or not. For this reason, the
contingencies of my epistemic life are not included in the aforementioned package of
needed parameters. They simply do not matter. In other terms, if the contexts are
different when I am rendered unconscious, then they are also different when I am not.
Epistemic states are surely important to the context of communication understood in a
broad and intuitive sense, but they have no significance for the semantically relevant
notion of context.

What seems to be happening in the example above is that I have mistaken
beliefs about the contexts in which (2) and (3) are uttered, and for this reason I take the
contents of “that” to be different. Could I also for this reason be mistaken about their
characters, taking them to be different when in fact they are not? This would explain why
their cognitive value differ to me. This line of reasoning, however, seems very doubtful. As
Kaplan says, “although a lack of knowledge about the context (…) may cause one to

---

mistake the Content of a given utterance, the Character of each well-formed expression is determined by the rules of language (...), which are presumably known to all competent speakers” (Kaplan, 1989a, p. 548). If this is right, then the difference in cognitive value in Taschek’s example arises not due to any incompetence or wrong beliefs regarding the characters of “that” in (2) and (3): I know all there is to know about their semantics (demonstration + demonstrative), and even so I mistake their contents. Put another way, I am not ignorant of any aspect of the meaning of (2) and (3); their characters are established independently of the contexts (and the referents) about which I have mistaken beliefs. Characters, remember, are given descriptively, and thus qualitatively, by linguistic norms (or, in the case of demonstrations, by the appearance of objects). Contrary to contents, characters do not directly involve or depend on any particulars about which misidentification is possible (and common): they are abstract types. If this is correct, mistakes about characters necessarily involve some sort of semantic incompetence, which is clearly not the case in the example above. Knowing the language is sufficient for knowing the character. Wrong beliefs about contexts, thus, cannot cause a speaker to be mistaken about the characters of the indexicals occurring in those contexts. Only some defect in linguistic competence can do so.

Contexts, then, are either not available to the referentialist sympathetic to the Fregean approach or cannot explain the phenomena even if its relevant parameters are somehow incorporated at the level of content. Furthermore, mistaken beliefs about contexts have no bearing at all on my competence with the relevant characters. Worse yet, the same kind of problem arises even when demonstrations of the same type are performed in the same context. Consider this example offered by Bozickovic (2009):

An illusionist may come up with a trick creating the impression that an object sitting in one spot has been replaced with another one qualitatively identical with it, whereas this is not so. In the process, he may say: “This is this”, where both utterances of ‘this’ are consecutively taken to refer to the same object, while relying on demonstrations that are of the same type (as in keeping his arm fixed in one position while uttering the sentence). (Bozickovic, 2009, p. 546).

In this case, the sentence “this is this” is uttered in a single context, and yet the cognitive values of the two occurrences of “this” are obviously distinct: I may think that the illusionist is saying something false or I may learn something new by hearing his utterance. Salmon (1986, pp. 74-4) has a similar example. Imagine that Marco has two
friends, Paul and Peter, who are identical twins, and hence qualitatively undistinguishable. Imagine also that Paul is standing in front of Marco. If I blindfold him for a few seconds, and then release the blindfold, Marco would not know for sure if it is Paul or Peter standing in front of him. If I say “he is he” during the entire process (the first “he” before the blindfolding and the second “he” after the blindfolding) I would speak informatively (or non-trivially). The two occurrences of “he”, therefore, have different cognitive values despite having identical characters. How could this be possible if in both cases the demonstratum, the demonstrative, the demonstration type and even the contexts are the same?

One could object that the contexts are not the same: they shift midsentence between the two occurrences of “this” or “he” in the examples above, and this is why their cognitive values differ. First, this is a very implausible view in its own right. As I argued earlier, being blindfolded of course effects a change in the context of communication understood in an intuitive and general sense, but it has no relevance of any kind to the appropriate notion of context, i.e., the package of parameters needed to supply the contents to expressions. Put another way, the illusionist’s trick and the blindfolding alter the subject’s awareness of the relevant context, not the context itself, which is the same for both occurrences of “this” and “he”. Moreover, consistency requires that we maintain that the context shifts in every occurrence of a demonstrative. This thesis, however, is far from obvious. Second, even if the contexts do shift, this does not explain why the same utterances can be trivial, as I argued above. If Marco is not blindfolded (or if the illusionist’s trick fails), the utterances of “he is he” (or “this is this”) should also be informative, for the contexts would be distinct for each occurrence of the demonstratives. But they would not be informative. This view generates informativeness where there is none.

Any appeal to contexts, then, seems doomed. A way of explaining those cases without resorting to contexts is by pointing out that the demonstration tokens are numerically distinct despite being of the same type. The cognitive value would then lie

---

42 Caplan (2003) apparently holds this view. Braun (1996) also discusses (but rejects) it. Perhaps there are some special words that can induce a context-shift midsentence, like “now” in “now is not now”. However, this is not the case in the given examples. In any case, the claim that contexts can shift midsentence is certainly not a very obvious one and it is something that must be argued for; it is certainly not in keeping with how we intuitively individuate contexts.
somehow in the numerical distinctness of the tokens: different demonstration tokens must have different characters. However, this is not what Kaplan’s theory says. Remember, according to Kaplan, a difference in cognitive value is explained by a difference in character. If demonstrations are types, then they are abstract things which cannot involve spatial or temporal properties, i.e., they cannot have any property that distinguishes one occurrence of the type from another occurrence of the same type. Thus, if the character is attached to the demonstration as a type, it cannot be affected by the circumstances of its tokening. All tokenings of the same type will share precisely the same character.

In order to solve this problem, it seems necessary to abandon the idea that character is applied to demonstrations as types. If it were possible to think of characters as being a property of tokens or occurrences, then it appears that characters could be sensitive to spatial and temporal properties, thus explaining the difference in cognitive value in terms of a difference in those properties. Different demonstration tokens would differ at least in temporal properties. If this is the case, then each tokening would generate different characters, which would entail different cognitive values, explaining the informativeness of the examples above.

This, however, starts to sound as a far cry from Kaplan’s original theory. First, this new notion of character as a property of tokens seems very different from the notion of character as linguistic meaning, in the sense of a general rule fixed by the conventions of language. Each demonstration token will have a new character, which is contextually determined every time a demonstrative reference occurs. In other words, the character of a complete demonstrative would be produced on the spot, and not established prior to its utterance (or occurrence). The character, therefore, will no longer be stable and fixed as Kaplan originally held, but generated only in the occasion of its production. The idea that characters are contextually determined seems incompatible with the idea that characters are linguistic meanings, which are evidently fixed by the rules of language prior to any utterances. Also, in order to preserve compositionality, characters cannot be applied to linguistic expressions as types either, because this would require the strange union of tokens and types I mentioned above. Thus, character should be a property of linguistic expressions as tokens (or occurrences).
Second, it does not sound wrong to say that a demonstrative, as a word, has a linguistic meaning which is constant and stable, like a dictionary definition. But what about demonstration tokens? If demonstration tokens have characters and characters are linguistic meanings, then it should not seem strange to say that demonstration tokens have linguistic meanings. But it does. Tokens are potentially infinite; it is not possible to provide dictionary definitions to all of them. Of course, demonstrations seem rule-governed such as linguistic expressions, but to say that these rules are *linguistic meanings* is to stretch this notion to the point of rupture. Moreover, even if we consider the rules that govern demonstrations as some sort of linguistic meaning, these rules would have to be very general; hence, they would not be fine-grained enough to individuate cognitive value, for the same rule would govern similar demonstrations that nevertheless have distinct cognitive value. We again run into the same problems that the above examples raised\(^43\). This modification in Kaplan’s theory, then, introduces a serious tension between the notion of character and of linguistic meaning.

What this reasoning suggests is that, if we want to solve the problem of cognitive value *via* the notion of character, i.e., *via* an aspect of the meaning of indexicals, we should resolve this tension by abandoning properties (ii) and (iii) of characters:

ii) It is the linguistic meaning of indexicals

iii) It is a property of expressions as types

Linguistic meanings then would be one thing, while character would be another. They would be different semantic features of the same expression. The former would still be a property of expressions as types and fixed by the norms of language, as Kaplan wanted, but the latter would be a property of their tokens (or occurrences) and would be determined only contextually, on the spot, when the demonstratives are associated with a demonstration token. This, it seems, could explain why two occurrences of the same complete demonstrative can have different cognitive values: the same word, when coupled with distinct demonstrations tokens, acquire distinct characters, and hence different cognitive values. Luckily enough, there are also *semantic* reasons for

---

\(^{43}\) Wettstein (1986) in his footnote 9 raised a somewhat similar problem with rule-governed demonstrations.
2.6 Braun’s revision of Kaplan’s theory

Braun’s (1996) argument is fairly simple. Forget Kaplanian semantics for a moment and think about the sentence “that is bigger than that”, where I point to two cars of different sizes. It seems clear that both occurrences of “that” have identical linguistic meanings. After all, it is the same word, occurring twice. This is what our pre-theoretical notion of linguistic meaning would tell us. Now enter Kaplanian semantics. Kaplan says that characters are (i) what determine the referent (content) in a context and (ii) the linguistic meanings of indexicals. If both occurrences of “that” have the same linguistic meaning, then they have the same character. If they have the same character in the same context, then they determine the same referent, and thus the same content. Hence, “that is bigger than that” would express the proposition that a certain object is bigger than itself, which is obviously a contradiction. Worse yet, this sentence would express this false proposition in every context in which it is uttered.

Braun notes, however, that in his formal language for indexicals Kaplan seems to avoid this problem. He would translate the above sentence somewhat as follows:

(a) That[d1] is bigger than that[d2]

This means that in his formal language for indexicals Kaplan seems to think that multiple occurrences of the same demonstrative must have different characters in order to refer to different objects: the function of the symbol [d subscript] is precisely to distinguish one character from another in order to secure distinct objects in the context. “That[d1]” refers to the first object in a sequence of objects, “that[d2]” refers to the second object in that sequence, and so on. If the characters are different, then two different objects are determined, thus ensuring that the proposition expressed is not a contradiction.

But what do the subscripts accompanying the demonstratives represent in natural language? The answer seems clear enough: demonstrations. In his formal language Kaplan apparently made a better characterization of demonstratives, that is, he effectively incorporated the character of the accompanying demonstration to the character of complete
demonstratives. The character of a complete demonstrative then is formed by combining a demonstrative with a demonstration.

What we need to make Kaplan’s theory and his formal language coherent, Braun argues, is to draw a sharp distinction between the linguistic meaning of an indexical and the character it has in the context of its use. Kaplan’s original two meaning theory of indexicals needs to be revised; we need a three meaning theory instead: linguistic meaning, character, and content. Linguistic meanings are what we intuitively think they are, i.e., the general rules of language attached to the words, or what I earlier called dictionary definitions. Together with the accompanying demonstration, linguistic meaning determines a character, which in turn determines the content. Putting it more formally, linguistic meanings of demonstratives are functions from demonstrations to characters, and characters are functions from contexts to contents. We then have this extended picture:

\[
\text{Linguistic meaning} + \text{demonstration} \rightarrow \text{character}
\]

\[
\text{Character} + \text{context} \rightarrow \text{content}
\]

To sum up. Character is the semantic feature of expressions that determines the content. If two characters are identical, then they determine the same object in the context. Thus, we need some distinction between the characters of the two occurrences of “that” in “that is bigger than that”; if we do not distinguish them, this sentence will always express a contradiction, for it will always refer to the same object twice. The correct way to solve this problem, Braun argues, is to distinguish linguistic meaning from character. The demonstrative by itself in this revised Kaplanian semantics has no character, only linguistic meaning. Linguistic meaning, in a context plus a demonstration, determines a character. It is this character that determines the content. In this way, Braun hopes to refine Kaplan’s theory, making clearer the importance of demonstrations to the determination of content and avoiding the complications arising from the identification of character with linguistic meaning.

It is worth noting that Braun’s revision does not imply that the associated demonstrations must be considered as types or that they must be considered as tokens. His view seems compatible with both views. In other words, for purposes of determining

---

the referent, it makes no difference to consider demonstrations as types or tokens in Braun’s three meaning theory. Both views account equally well for the problematic cases he analyzed.

However, for the reasons discussed above, if we want to explain Frege’s Puzzle resorting to demonstrations, then we have to consider them as tokens. To illustrate this, let us look at Braun’s picture, interpreted as taking demonstrations as types:

\[
\text{Linguistic meaning + demonstration (type) } \rightarrow \text{ character}
\]

In Taschek’s example, it is clear that both occurrences of “this” will have the same character if the accompanying demonstrations are considered as types. As we saw, both demonstrations exploit exactly the same perspective, appearance, gesture, etc. They are identical in every aspect except for the time of tokening, which is irrelevant to the character as a property of types. So, both utterances of “this is an F” will involve the same demonstration type, thus generating the same character two times. From a semantic point of view this is harmless. In both utterances the same proposition is being expressed, and this fact is predicted and explained by this interpretation. However, if we want characters to explain cognitive value, the characters that result from the combination of linguistic meaning + demonstration must be different. This is achieved by considering demonstrations as tokens. We will then have the following picture:

- **Utterance 1**: Linguistic meaning + demonstration (token₁) → character₁
- **Utterance 2**: Linguistic meaning + demonstration (token₂) → character₂

### 2.7 Kaplan’s revised theory and the solution (?) to Frege’s Puzzle

We saw that we have some reasons, both epistemic and semantic, for distinguishing character from linguistic meaning. On the one hand, there is the linguistic meaning of expressions, which is constant and it is fixed by convention. On the other hand, there is the character, which is determined by linguistic meaning plus the associated demonstration and has the function of determining the content in a context. It is the latter feature that accounts for phenomena of cognitive value. Consider Taschek’s example. Assuming that demonstrations are tokens, each performance of a demonstrative reference will necessarily involve distinct demonstration tokens. If the demonstration tokens are
distinct, then they will determine different characters for each utterance of the demonstrative “this”. Hence, no two occurrences of the same demonstrative will have the same character. This is why (2) and (3) have distinct cognitive values: despite involving the same demonstrative (word), they have distinct characters.

The same explanation goes for any counter examples like those offered by Wettstein and Recanati. Again, no two demonstrative references will ever have the same character, for they will always involve numerically distinct demonstration tokens. This is why accepting the sentence “this ship$_1$ is a steamer but this ship$_2$ is not a steamer” does not make me irrational or why “he$_1$ is he$_2$” can be informative and non-trivial. The content is precisely the same, but it is presented to my intellect under different modes of presentation because two distinct demonstration tokens are at play. The character, which is generated on the spot, individuates cognitive value after all.

Despite the foregoing, this solution generates a problem that is, to my view, insuperable. It explains nicely the informativeness and the different cognitive values that demonstrative references may have, but what about their triviality? Think of a trivial case of “this is this”, where I point twice at the same object. Each pointing will generate a different character. For purposes of the determination of reference, this makes no difference. Different characters can determine different objects in a context, but is not necessarily so. However, if different characters entail different cognitive values, how can a sentence involving distinct characters nevertheless be trivial? We can think of innumerable examples like this one. There are trivial cases of “he is he”, of “that is that”, of “that is this”, etc. They will involve distinct demonstration tokens and so they will have distinct characters, but they are trivial nonetheless. If different characters determine different cognitive values, then we would have informativeness where in fact there is none.

To summarize. We saw that linguistic meaning is too coarse-grained to individuate cognitive value, for there are sentences that involve precisely the same linguistic meaning and are nevertheless informative. We tried to solve this problem by stressing the importance of demonstrations in demonstrative reference. We noted that demonstrations are semantically relevant to determining the content in the case of demonstratives, so they can affect the character of a complete demonstrative. However, if demonstrations are considered as types, they are incapable of explaining examples like
Taschek’s, where the same demonstration type is involved twice and yet there is a
difference in cognitive value. Character must then be a property of tokens, so it must be a
semantic property that is determined on the occasion of the tokening. There are also good
semantic reasons for this thesis. This apparently solves the problem, because two
numerically distinct demonstrative references will never involve the same demonstration
token twice, so they will always generate distinct characters, which entail distinct cognitive
values. But this creates a problem for explaining triviality. If character is what explains
cognitive value, then different characters should entail different cognitive value. But this is
not the case in trivial examples of “he is he” or “this is this”. The characters of the
demonstrative references are clearly distinct, and yet those sentences are trivial and
uninformative. This amended Kaplanian semantics solves one problem but creates another
at the same time.

### 2.8 A different solution

One can attempt to rescue Kaplan’s solution to the problem of cognitive value
by making a move that can be derived from Kaplan himself. He said that a single ostension
can have many different demonstrations associated with it: a single gesture can produce
many different perspectives, and each perspective would count as a different
demonstration\(^{45}\). In other words, Kaplan believed that a single demonstrative performance
can produce distinct manners of presentation depending on the location of the audience.
For example, if I point at something, those to my left will have one cognitive perspective
over the referent, while those to my right will have a different perspective. Hence, the same
act performed by the speaker can produce many different cognitive values for the audience.
At the same time, there seems to be nothing to prevent distinct demonstrative gestures to
produce the same mode of presentation. I can point twice at the same object, but my
pointing would generate identical perspectives. The consequence of this view is that
cognitive value seems to be something much more relative, because it depends heavily on
the side of the audience and not so much on the side of the speaker. Hence, the same
demonstrative reference may potentially produce many different cognitive values, while
distinct ostensions can produce the same cognitive values.

This idea can be interpreted in two ways: (1) the character is exactly the same for every occurrence of a demonstrative, but the accompanying demonstration can provide many different cognitive perspectives over the referent, thus explaining the informativeness or the triviality of sentences; this being so, the cognitive value is affected by the demonstration, but the character is not; (2) there is no single character determined by the speaker alone, but each possible cognitive perspective is in fact a different character; hence, the character depends on the audience and on their perspectives over the referent, not on the speaker who performed the demonstration.

The first problem with interpretation (1) is obvious. If we assume that cognitive value is explained by character, as Kaplan does, then this interpretation is wrong from the start. Under this assumption, the character cannot remain the same while producing different cognitive values. This was the problem that motivated the revision in Kaplan’s theory in the first place. On the other hand, if we drop this assumption, then there is no hope for solving Frege’s Puzzle for indexicals in terms of meaning. We can all agree that cognitive value is affected by the associated demonstration. Yet, if demonstrations do not affect the character of the expression, then we can hardly say that they are semantically relevant. We say that the character is semantically relevant because it is the feature of expressions that determines the referent. In other words, it performs a role in determining the propositional content of the sentence, in establishing its truth conditions. By contrast, if the function of demonstrations is simply to provide modes of presentation of the referents and not to help determining them, then their function lies outside the sphere of semantics. The relevance of demonstrations is merely epistemic. This of course is not a problem for the semanticists who do not accept the Fregean model of semantics, but it is a huge blow to those who do, for they need to solve Frege’s Puzzle in terms of the semantic properties of expressions, and not in terms of something that has no role in semantics whatsoever.

The interpretation (2) preserves the assumption that cognitive value is explained by the character, but runs counter a very plausible intuition about the functioning of language: when a demonstrative reference is made, it is the speaker who determines a single character that must be grasped by the audience, not the audience who determines number of characters according to their perspectives over the referent. Let me explain.
When a reference is made, we intuitively say that the speaker is responsible for correctly using and appropriately the conventions of language. The role of the hearer in a communication exchange is to interpret the actions performed by the speaker in the appropriate way. Some examples can be provided to support this idea. Suppose, for instance, that there is just one woman in front of me and my friend and I say “she is really gorgeous” pointing right at her. Suppose also that my friend, for some reason, is not really paying attention to me or that he is staring at another woman behind us. I think that it is plausible enough to say that I was successful in determining a referent: I used the correct indexical and the correct demonstration unambiguously, thus exploring accurately the right conventions for demonstrative reference. The fact that my friend did not grasp what I said seems irrelevant to the success of my reference. This is even clearer if we imagine that, due to my bad timing, the woman heard what I said and got really upset. It seems strange to say that for my friend the reference was not successful while for the woman it was successful. I intended to convey some information to him, not to the woman, but even so she was able to grasp what I wanted to say. In fact, we can say that had other people been there, they would have understood me correctly. This statement would not make any sense if reference was not something done solely by the speaker. The success or failure of reference, it seems, is a matter of exploiting accurately the rules established by the linguistic community, and not a matter of divided labor between speaker and audience. Understanding a demonstrative reference is a whole other matter.

Other example that can be offered to support the idea that reference is a matter of correct using linguistic rules by the speaker is this. Imagine that I am giving a lecture somewhere where there used to be a very impressive painting on the wall just behind the place where I am standing. However, this painting was stolen days before and the university replaced it with a much less remarkable statue. Being very absent-minded, I did not notice the replacement when I started lecturing. At some point I say, pointing with my thumb over my shoulder, “that is a remarkable painting”. The audience laughs and I turn to see why. Surprised with what I just said, I say “Sorry, what I said is absurd. What I meant to say is that the painting that once hanged there was remarkable”. What can we say about this case? It seems clear that I exploited perfectly all the conventions for demonstrative reference, but nevertheless ended up saying something I did not want to say. I pointed to a

---

46 This example is borrowed and adapted from Kaplan (1978).
single object, predicated something of it and was impeccably understood by the audience. I obviously said something false, because the thing I pointed at is not even a painting. But *had it been the painting I had in mind*, I would have said something true. The fact that I intended to refer to other object has no relevance whatsoever to what I in fact said. Indeed, I would not even correct myself by saying “No, you got me wrong, *I said* that the painting that once hanged there was remarkable”. I did not say that. Even assuming that the audience is very comprehensive and have interpreted me in the most charitable way, this would not change the fact that I referred to the statue, not to the painting. The rules I exploited to make this reference dictate that the proposition expressed has whatever object I pointed to as a constituent. What I in fact said may be a surprise even to myself.

Other examples could be given, but I think those are enough. The upshot of this reasoning is that reference seems to depend on the correct use of the rules of language, which are pretty much independent of possible hearers and of our own general intentions. In other words, the linguistic conventions determine what we say, independently of our intentions. Our ability to make successful references depend on our ability to use those rules in the right way. Thus, it seems that the job of securing a referent is attributed only to the speaker, and not somehow to both speakers and hearers. Hence, if it is the speaker alone who refers, then it *is he alone that determines the character*, not the possible hearers. The task of the audience is to interpret the single character the speaker expressed, so that it can grasp the content he conveyed. In other words, interpretation (2) says that innumerable characters can be generated by a single utterance, but this goes against the idea that it is the speaker’s responsibility to secure a referent. Moreover, communication seems very mysterious if we allow character to be multiplied in this way. What makes speakers understand each other if the semantic feature that guides them to the referent can be completely different for each one of them? What makes you understand *what I said* if what determines the referent is something radically distinct for you and me? The character must

---

47 It is worth noting that general referential intentions need not be *directing intentions* in Kaplan’s (1989b) sense. In this example, I had the general intention to refer to the painting, but my *directing intention* was towards whatever was behind me, which unfortunately happened to be a statue. Similarly, I may have the general intention to refer to Frege by pointing at his picture, but my *directing intention* seems to be towards the picture itself. I can refer to Frege in virtue of referring to the picture. Anyway, directing intentions are a complicated and controversial matter, and I do not want to commit myself with any particular version of it. I believe, however, that directing intentions are inessential to the case I am making here. It is enough if my arguments apply to general referential intentions only. For arguments in favor of directing intentions, see Kaplan (1989b) and Perry (2009). For very compelling arguments against directing intentions, see Reimer (1991a) and (1991b).
be univocal for both speakers and hearers. Interpretation (2) makes communication less a social enterprise and more an individualistic guessing game.

The idea presented at the beginning of this section, then, is mistaken. If we detach demonstrations from characters, then they will not have any semantic relevance and thus cannot explain cognitive value on semantic grounds. On the other hand, if we allow characters to be multiplied indefinitely by the number of hearers, then we arrive at a very implausible picture of how language works.

There is one more solution available, offered by Corazza and Dokic (1992). It is not exactly a Kaplanian solution, for they do not seem to think that the character is capable of accounting for Frege’s Puzzle, but it is close enough and they agree with the general Fregean approach to semantics. However, I think their proposal suffers from serious problems that cannot be avoided.

2.9 Corazza and Dokic: acts of perception

Corazza and Dokic (1992) accept Kaplan’s distinction between demonstratives and pure indexicals, i.e., the distinction between indexicals that require some sort of complement and indexicals that do not. However, they say that demonstratives are to be combined with acts of perception, not with pointing gestures or with demonstrations in the broad sense, as Kaplan suggested. They do not define precisely what an act of perception is, but the idea is sufficiently intuitive to be understood and worked with. They say, for instance, that Evan’s notion of keeping track of an object is compatible with their notion of an act of perception (although they do not explain exactly how).

But how this solves Frege’s Puzzle? The idea is simple. An instance of “he is he” is trivial if it is based on a single act of perception, while an instance of “he is he” is informative if it is based on two distinct acts of perception. If I point to an object twice without looking away, for example, the demonstratives used will be completed with the same perceptual act; hence, my utterance will be uninformative. However, as in Recanati’s ship example, the two occurrences of “this ship” are based on distinct perceptual acts, because the ship is obstructed in such a way that is not possible to track it continuously along its whole extension. In other words, two perception tokens are employed in the utterance of “this ship₁ is a steamer but this ship₂ is not a steamer”. Hence, the two
occurrences of “this ship” have distinct cognitive values despite having the same linguistic meaning. Of course, Corazza and Dokic are distancing themselves from Kaplan, because this solution does not assume that the character is what accounts for cognitive value. They do, however, conform to the Fregean model of semantics, for acts of perception are, on their view, semantically relevant to determining the referent.

However ingenious this solution may be, I believe it does not work for the following reasons. It is not very clear how it is possible for perceptual acts to have semantic relevance. They might explain why and how we understand demonstrative reference, but it seems to me that they cannot perform any function whatsoever in determining the reference of an utterance. At best, perceptual acts distinguish and determine one object for the speaker, but his perceptions cannot be conveyed by any linguistic devices, so they are irrelevant for purposes of determining the semantic content of an utterance. In other words, perceptual acts might determine for the speaker’s cognition the object he intends to refer to; however, since he needs to go from private perceptions to language, to communication, he must exploit public conventions that are shared by the linguistic community in order to secure the referent.

Worse yet, perceptions are not always involved in securing a referent even for the speaker. Think of the lecture example I gave before. I did not perceive the object I was demonstratively referring to – the statue behind me – and yet my reference was successful. Of course, I relied on previous perceptions to do that because I believed that there used to be such and such a painting behind me, but an actual perception did not play any role in my actual demonstrative reference. It was not perception that determined the referent, but the fact that I pointed at a single object and exploited all the correct conventions of demonstrative reference. On the side of the audience, perception is needed in order to fully understand what I expressed, but the referent was already determined by the rules and conventions of language I used. Perception needs to come into play if we want to explain how demonstrative reference is apprehended and fully understood, but this explanation is not part of semantics per se.

See McGinn (1981) for many other examples of demonstrative reference that do not involve perception of the demonstratum.
Pointing gestures are clearly capable of discriminating one object from another in a context. They exploit *public* rules and conventions that are shared by the whole linguistic community, but the same thing cannot be said about acts of perception. Perceptions are essentially private things. How can they be *semantically* relevant in a communication exchange? I cannot rely on *my* perceptions in order to secure a referent for the audience. If this were the case, communication would be a solipsistic exercise.

Even if we grant that perceptual acts *do* have semantic relevance, they would either be unable to explain cognitive value or would incur in the same implausible view of communication explained in the previous section. Think of Recanati’s ship example. For me, there are two acts of perception involved, so each occurrence of “this” in “this ship1 is not this ship2” will have different cognitive values. If we hold that it is the speaker’s responsibility to secure a referent, and that acts of perception are semantically relevant, then only *my* acts of perception are responsible for endowing the utterances of “this” with the semantic properties they have, and hence with the cognitive value they have. Since the two occurrences of “this ship” have distinct cognitive values to me, they should also have distinct cognitive values to everyone else. However, imagine that I utter this sentence to my friend Clark Kent, who unbeknownst to me has X-ray vision. He can see through the obstruction perfectly. It seems obvious that the two occurrences of “this” will have *exactly the same* cognitive value for him, and he will know right away that I am expressing a contradiction, despite the fact that they have the same semantic properties. To solve this problem, we could argue that acts of perception are semantically relevant *for every possible subject* in the context, so the same utterance will have as many different semantic properties as the number of acts of perception involved, and thus as many different cognitive values as these acts of perception. In short, every listener determines the referent in her own way according to her own acts of perception. Needless to say, this is a highly implausible view of how communication occurs, as is the view that characters are multiplied according to possible differences in perspective over the referent.

It is possible to reply that what determines reference is in fact the demonstration performed, but the accompanying acts of perception are what explain cognitive value. This may very well be true, but again, if perceptual acts do not help to determine the referent in any way, then they are outside the scope of semantics. They are not an aspect of *meaning*. Perceptual acts then are not a semantic property of expressions.
in any plausible sense: they are an epistemological notion that explains cognitive phenomenata related to language, but they do not do any semantic work. If this is true, Corazza and Dokic cannot explain cognitive value from within semantics, as the Fregean approach requires.

2.10 Final blow to the solution via character: pure indexicals

So far we have been concerned with the problem of cognitive value as it appeared in sentences involving demonstratives. However, a more serious problem for trying to solve Frege’s Puzzle via character is posed by pure indexicals. Suppose I am drunk and happy at a party with my friends. One of them says to me “it is nice here” and I agree with him. Not much later, incredibly drunk and completely disoriented, I hear my friend saying “it is nice here” again at the exact same place and with exactly the same surroundings. However, wrongly believing I have moved to another place, I do not agree with him. Those two utterances obviously have different cognitive value to me, because I accept one while denying the other. It is important to note that I have not changed my mind regarding the one place in question: I still believe what I believed when my friend first uttered “it is nice here”, but I disagree with the second utterance because I think I am at a different place. Am I irrational? I think not. How can Kaplan’s theory deal with this?

Even if the character (or acts of perception) explained successfully cognitive value in the case of demonstratives (which it did not), it would not be able to account for this case. This is because it is not possible to resort to demonstrations in order to explain the difference in cognitive value between the two utterances of “it is nice here” as in the example above. The indexical “here” is strictly not a demonstrative; it is a pure indexical, which does not require a demonstration in order to secure a referent. The character seems to be precisely the same in both occurrences and no demonstration is needed for the indexical to secure its referent. In other words, this kind of indexical, i.e., pure indexicals, perform their semantic function of determining a referent in a context by themselves. There is no need for an accompanying demonstration of any sort, as we saw in section 1.1. The character of “here”, “now”, “today” or “I” does not require a complement to discriminate one single object in a context. Hence, its character is truly invariable in every possible

---

49 Caplan (2003), footnote 26 offers a similar example to make the same point.
context of their use. If the character is invariable, so should be the cognitive value. But this is clearly not the case.

Even if one claims that characters of pure indexicals are somehow distinct in every occurrence, i.e., generated on the spot, this explanation would face the same problem as the explanation for demonstratives. It accounts for informativeness, but not for triviality: because two distinct occurrences of a pure indexical would *always* differ in character they should also *always* differ in cognitive value; yet, as we saw, this is not the case.

Remember also that the semanticist sympathetic to the Fregean approach cannot resort to differences in the context or mistaken beliefs about the context to explain differences in cognitive value, as I argued earlier. It seems obvious that these things play an essential role in determining the cognitive value of the two utterances in question, but contexts are not semantic properties of expressions. They are not in any way a feature of meaning. Furthermore, the character is independent of the context; no mistaken belief about the context can affect the speaker’s knowledge about the character.

To recapitulate, then. The only way to make the character of demonstratives able to explain cognitive value is to consider the demonstrations that complete demonstratives as tokens, not types. However, this move generates a difference in cognitive value where there is none, because no two demonstrative references will involve the same demonstration tokens, and hence no two demonstrative references will have the same character. On the other hand, Corazza and Dokic’s solution fails as well, because it relies on the notion of acts of perception, which does not seem to perform any semantic work. If it does not, then this apparent explanation of Frege’s Puzzle is not given in terms of semantic properties of expressions, but rather by appealing to concepts outside the sphere of semantics. This gets even more complicated because Frege’s Puzzle can arise for pure indexicals too. In this case, there is not the possibility of appealing to demonstrations or acts of perception, because pure indexicals do not need to be complemented in any sort of way. Just their utterance is sufficient to determining the content in a context, and hence no other semantic feature is needed.

So, it seems that the solution to Frege’s Puzzle in the case of indexicals cannot be given in terms of the character of those indexicals. Every attempt to do that seems either to slip outside the sphere of semantics, thus not adequating to the Fregean model, or simply
to fail to account satisfactorily for all phenomena of cognitive value. Yet, not all hope is lost for the direct reference theorist who wants to explain cognitive value as a property of meaning.

2.11 The solution via reflexive content and concluding remarks

In light of Wettstein’s (1986) criticisms, Perry (1988) offered a different solution to Frege’s Puzzle as arising for indexicals in terms of what he called reflexive content. Perry (1988) distinguished between the proposition expressed by an utterance and the proposition created by an utterance, and argued that the latter is what explains cognitive value. The proposition expressed by an utterance \( u \) is its official content (or content\(_C \), in later terminology)\(^{50} \), i.e., what we would normally regard as what is said in the context, its referential truth conditions. On the other hand, the proposition created (or content\(_M \)^{51} ) by an utterance \( u \) is the proposition which is generated on the occasion of its production, which has the utterance \( u \) itself as constituent. This created proposition, which Perry calls the reflexive content, states the conditions under which the utterance \( u \) is true. This kind of propositional content is determined solely by the linguistic meaning associated with the sentence-type of which utterance \( u \) is a token. Put another way, the reflexive content is a product of the linguistic rules which are attached to linguistic expressions as types and of the occasion of their utterances. These linguistic rules can be thought of as the characters of the relevant expressions. Note also that the speaker does not need to explicitly and consciously think about the reflexive content of an utterance; since this content is derived from linguistic rules, it is presumably grasped automatically and effortlessly by every competent speaker of the language.

Let us see an example. Consider a situation where, pointing to a certain professor, I utter to you “she is brilliant”. According to Perry, to fully account for what goes on in this situation, we must distinguish between various levels of meaning. First, there is the character of “she” and of “is brilliant”, which is associated to these expressions as types by the linguistic norms. Second, there is the proposition I literally express in this context, which is the official content (content\(_C \)), of my utterance: the singular proposition containing the professor and the property of being brilliant. Third, there is the reflexive

\(^{50} \) Cf. Perry (1997), p. 11.
\(^{51} \) Idem.
content \( (\text{content}_M) \), which is something along the lines of “there is one discriminated female \( x \) which is the referent of \textit{this utterance} of ‘she is brilliant’, and \( x \) is brilliant”\(^{52}\). The same goes for every indexical that is part of an utterance of a complete sentence. For example, the reflexive content of an utterance of “you” will be “the addressee of \textit{this utterance} of ‘you’”, the reflexive content of an utterance of “here” will be “the place which is referred to by \textit{this utterance} of ‘here’”, and so on. As we can see, the reflexive content does not contain the demonstratum as an ingredient; it contains only the utterance \( u \) itself and a general description of the demonstratum in terms of the context and its relation to the utterance \( u \).

This is why one can in some sense understand \textit{every utterance} of “she is brilliant” even if one does not know the referent of “she” in the context: every utterance \( u \) of this sentence-type creates a proposition that states the conditions that must be met in order for the utterance \( u \) to be true in the relevant context, and these conditions are derived from the linguistic meaning – the character – of the relevant expressions. Thus, the hearer grasps the created proposition even if she does not grasp the official content expressed by the utterance, which requires more than just linguistic knowledge to be grasped. More importantly, since the reflexive content of an utterance has \textit{the utterance itself} (and its relevant parts) as a constituent, each new utterance will produce a different reflexive content, i.e., a different created proposition.

This is why, according to Perry, the utterances of “it is nice here” or “that is an F” differ in cognitive value: although their official content is the same, the propositions they create are distinct. Since the reflexive content describes the demonstratum in terms of its relation to the context and to the utterances themselves, and since the utterances are different, the created propositions will describe the demonstratum by expressing different relations, and hence will provide different cognitive perspectives over it. Thus, a difference in cognitive perspective corresponds to a difference in cognitive value. As Corazza and Dokic put it, although linguistic meanings, understood as types, do not determine directly the cognitive value, they do so when \textit{applied} to a given utterance\(^{53}\). Perry’s reflexive content, since it is clearly an aspect of (applied) meaning, thus respects the Fregean criterion of adequacy for semantics.

Given our previous discussion of character and demonstrations, the problem with Perry’s solution may seem apparent by now. If we individuate cognitive value in terms of utterances, numerically different utterances should have different cognitive values. However, this is clearly wrong. We have many trivial instances of “it is nice here” or of “that is an F”, as well as trivial utterances of “this is this”, “that ship is that ship”, “she is she”, etc., even though each created proposition will have different constituents. In other terms, every utterance of an indexical will contribute itself to the propositions that are the reflexive contents of the relevant utterances; hence, these utterances should always differ in cognitive value. But this is obviously not the case. Utterances are unrepeatable and always numerically distinct, as are demonstrations tokens (or occurrences) in Kaplan’s revised theory. It is not surprising that they have the exact same problem: they are simply too fine-grained to individuate cognitive value correctly.

If the foregoing discussion is correct, the prospects for the referentialist who wants to solve Frege’s Puzzle on semantic grounds look gloomy. Neither the character (even in its modified version) nor the reflexive content seem able to explain adequately all phenomena of cognitive value for indexicals. This is especially worrisome because indexicals, as I argued earlier, are the type of expression that seems most hospitable to a semantic theory that conforms to the Fregean criterion of adequacy: their functioning in language naturally suggests that they have more than one level of meaning, one of which could possibly play the epistemic role of Fregean senses without also having to play their semantic role. However, as we saw, none of these levels of meaning are capable of solving Frege’s Puzzle, at least in its indexical version. It now remains to be seen how well character and reflexive content do when it comes to proper names (spoiler: not very well).

54 Corazza & Dokic (1992) and Loeffler (2001) make the same point about Perry’s reflexive content.
3. Proper Names

3.1 The character of names

As we saw in the first chapter, Frege’s Puzzle appears not only for indexicals, but for singular terms across the board. Names, then, are no exception to the conundrums of cognitive value. But they pose a special difficulty for referentialist who accepts the Fregean criterion of adequacy for semantics. Even though the two purported solutions to the indexical version of the puzzle examined in the last chapter ultimately failed, at least that version was more amenable to a meaning-related treatment given the natural distinction between the various levels of meaning of indexicals. Frege’s Puzzle for names, however, is nowhere near the amenability of its indexical version to a solution on semantic grounds. The reason is simple: for most direct reference theorists (Kaplan and Perry included), names are just labels for their referents. There is no other level of meaning to a proper name other than the object it stands for. In short, the meaning of a proper name is exhausted by its referent. If this is right, then all coreferential names, no matter how syntactically distinct, have exactly the same meaning. Or, as Kaplan puts it, in proper names “all three kinds of meaning – referent, content, and character – collapse. (…) Because of the collapse of character, content, and referent, it is not unnatural to say of proper names that they have no meaning other than their referent” (Kaplan, 1989a, p. 562). Therefore, there is no difference in meaning between “Spider-Man” and “Peter Parker”, or between “Hesperus” and “Phosphorus”, that can be exploited by the referentialist in order to account for the informativeness of “Spider-Man is Peter Parker” or “Hesperus is Phosphorus”. All there is to the semantics of coreferential names is simply identical. How can the referentialist account for the cognitive value of proper names on the Fregean approach?

Kaplan himself acknowledges the difficulty: “The problem is that proper names do not seem to fit into the whole semantical and epistemological scheme as I have developed it” (Kaplan, 1989a, p. 562). Kaplan, then, cannot solve Frege’s Puzzle as arising for names. His theory does not have enough resources to do so. In fact, this is a reason to suspect in principle his attempt to explain cognitive value via character, even in the case of indexicals. Let me explain. If we have a certain class of phenomena and a theoretical entity that purports to explain them, then the fact that this entity does not even begin to explain a
recognized subclass of the same phenomena gives us enough reason to suspect that it was not the entity that we were looking for in the first place. Therefore, if character, understood as a semantic property of language in general, really has the epistemic dimension that Kaplan claims it has, it should have this dimension for all singular terms, not only for indexicals. In other words: if cognitive value could plausibly be explained by character, then all phenomena of cognitive value should at least be initially treatable via character. But, in the case of proper names, they clearly are not. Characters seem ill suited to explain the cognitive value of names right from the outset. If this is correct, then it looks like that the apparent relation between cognitive value and character in the case of indexicals that Kaplan was so enthusiastic about was merely incidental. Kaplan, maybe impressed with his theory of indexicals, overestimated the explanatory power of the entities it postulated.

But let us not be so pessimistic. Perhaps Kaplan is wrong regarding proper names. Perhaps they do have more than one level of meaning besides the referent. If the referentialist argues that character and content in proper names do not coincide as Kaplan believes, then maybe the character is able solve the name version of the puzzle after all. This is obviously a major deviation from Kaplan’s original theory. But, as we saw, if the referentialist does not part ways with Kaplan regarding proper names, then she has no hope for solving the name version of the puzzle in terms of character.

There are several ways of explaining what character for proper names would be like, but all of them necessarily involve descriptions. There is no other way of doing that: characters are rules necessarily given by certain descriptions. Hence, the content of a proper name in this modified Kaplan’s theory would still be its referent, but its character would be some sort of description. It is important to stress that this is not in any way incompatible with direct reference. The biggest lesson from direct reference is that names are not equivalent to descriptions at the level of content. It is perfectly compatible with the general direct reference framework, then, for proper names to have a level of descriptive meaning very much like indexicals (direct reference also gave us very good reasons to suppose that proper names do not have descriptive meanings at all, but let us put this skepticism aside for a moment). The descriptions that state the characters of proper names
just have to function like descriptions coupled with Kaplan’s *dthat* operator\(^{55}\): they express conditions that must be satisfied by an object in order for it to be their extension, but they contribute only their extension to propositional content. In short, the *dthat* operator turns descriptions into directly referential terms. The character of a proper name, then, could very well function as a description of its bearer combined with a *dthat* operator\(^{56}\).

However, contrary to indexicals, which have more or less easily statable descriptive characters, finding which descriptions are good candidates for being the characters of proper names is a task of enormous difficulty. Since they function as characters, these descriptions have to be somehow cognitively accessible to all speakers who are competent with the name, and they must be responsible for determining its referent. The options for the characters of proper names seem to be the following:

1. Causal-historical chain description
2. Specific-name metalinguistic description
3. Generic-name metalinguistic description
4. Context-sensitive description

Options (1) and (2) are not so distant from Kaplan’s original theory, for they treat names as non-indexicals, i.e., as expressions whose character is constant and whose content does not depend on certain parameters of the context. Options (3) and (4), on the other hand, treat proper names as indexicals, i.e., as expressions whose character is context-sensitive. Let us begin by examining option (1).

Many people, including Kaplan, were convinced by Kripke’s causal-historical picture of how names refer. According to Kripke, a given tokening of a name refers to the object it does because it is part of chain that goes back to an initial baptism, when the name was first introduced as a name of the baptized object. This name is passed on from speaker to speaker, and it is in virtue of being causally connected to the object itself that my utterance of that name refers to the precise object it does. In short, the reference is fixed externally, by the name’s ancestry, not by some fact internal to my cognition.

\(^{55}\)Cf. Kaplan (1989a), p. 521-22, and Kaplan (1989b), p. 578-82. Kaplan says that *dthat* should be understood as a true demonstrative, not as a real operator. For our purposes here, however, this does not make much difference.

\(^{56}\)Note that the *dthat* operator is not the same as the *actually* operator, which is usually appealed to by descriptivists about proper names to deal with Kripke’s objections. The former stays at the level of character, so to speak, while the latter carries over to content. In other words, the *actually* operator just rigidifies a description; it does not turn it into a directly referential term as *dthat*. 
It is also important to stress that, in this picture, names are individuated in terms of baptism ceremonies. So, two coreferential names are the same name if and only if they were introduced by the same ceremony. Conversely, two coreferential names are different if and only if they were introduced in the linguistic community by two different baptisms. Hence, to each name corresponds one and only one causal-historical chain (more like a tree, actually), which is originated at a baptism ceremony. A consequence of this view is that, for example, a noun like “Ludovic” will be systematically ambiguous: there is not a single name “Ludovic”, but as many different names – spelled identically – as there are baptism ceremonies. In fact, if the same person is named “Ludovic” by two distinct ceremonies, there will be two distinct names spelled “Ludovic”, not a single one.

Kaplan believes that the role of this causal chain is pre-semantic or, as he puts it, metasemantic. This means that the causal chain is not somehow built into the meaning of a given proper name; it functions only to determine which name is being used, and hence which thing is being referred to. So, at a given occasion of discourse in which a name is used, contextual cues determine which causal chain is being exploited, and thus which name is being tokened, in the same way as contextual cues determine the meaning of an ambiguous expression such as “bank”. This is completely different from the way context determines the content of an indexical expression: for indexicals, the context-sensitivity is built into their characters, and thus into their meanings, whereas names have a context-insensitive meaning. In other terms, the referent of a name is not determined in virtue of an aspect of its meaning, but pre-semantically, by the chain that brought the name to the speaker. Context is relevant only to determine which chain is being exploited. It does not include a parameter which the meaning of the name is sensitive to. In short, to Kaplan, causal chains merely fix the referent in Kripke’s sense; they are not encoded in the name’s meaning.

Option (1) is the view that, contrary to Kaplan, causal chains are in fact encoded in the name’s meaning. More precisely, they are encoded at the level of character. So, the character of a name “N” would be given by a description such as “[dthat] the individual who lies at the other end of the historical chain that brought this token of ‘N’ to me”. As Kaplan puts it, such a theory will “regard the historical chain theory as a part of

---

57 This is essentially Kaplan’s example of the mischievous Babylonian. Cf. Kaplan (1990).
semantics, as giving the meaning rather than as telling us how to discover it” (Kaplan, 1989b, p. 574). If this is plausible, then names would have two layers of meaning: the character, which is given by a description of the causal chain that introduced the name in the community, and the content, which is just its referent.

This is why “Spider-man is Peter Parker” and “Hesperus is Phosphorus” are informative sentences. Since those names have clearly distinct causal histories, they have distinct characters. If they have distinct characters, and character is tied to cognitive value, then they have different cognitive values. Voilà: “Spider-man is Peter Parker” and “Hesperus is Phosphorus” turn out to be informative despite having the same referent.

There are mainly two problems with this solution. The first has nothing to do with Frege’s Puzzle, however. It is about the supposed cognitive role that causal chain descriptions can play. The characters of indexicals are more or less easily statable because they are rules of language that are known by the speakers. Put another way, the characters of indexicals are more intuitively conceived as linguistic meanings (or at least something that is derived from them) because they are, at least in some sense, grasped by every competent speaker of the language. And the descriptions that state these characters are fairly simple. Causal descriptions, on the other hand, are much more complex and seem much more cognitively demanding. It is certainly not very intuitive to say that causal descriptions are rules that have to be mastered by competent speakers for the correct use of names: they require substantive beliefs about baptisms, causal connections, linguistic communities, etc. These sorts of beliefs are obviously required for the linguistic practice in general. But so are beliefs about sounds, symbols, perceptual encounters, behaviors, etc., that are not built into meanings. The idea that the character of a name is a description of its causal chain, then, is very doubtful. Of course, this is far from being a knock-down argument against this proposal, but this view is certainly not in keeping with our intuition about what linguistic meanings are and the cognitive role they play.

Second, and more seriously, this proposal does not solve Frege’s Puzzle for names even if it is correct. Consider this situation described by Kaplan: “I may introduce a new proper name word and send it on its journey. When it returns to me – perhaps slightly distorted phonologically by its trip through other dialects – I can competently take into my vocabulary without recognizing it as the very same word! Shocking!” (Kaplan, 1989a, p. 563). If the view described above is right, in this case there is just one baptism ceremony,
so there is just one name with a single character, call it ‘N’. When I encounter this name again, it may be nevertheless informative to me to be told that “N is N”, even though I was the one who introduced it in the first place. The name does not even have to be phonologically distorted; it can be spelled and pronounced in the exact same way as I introduced it. This also occurs in Paderewski-like scenarios. It may be informative to me to be told that “Paderewski is Paderewski” even though both occurrences of the name exploit the same causal chain, and thus have the same character. How can this be possible?

Option (1), then, seems barely tenable as a solution to Frege’s Puzzle. Option (2) seems a little bit more plausible, but it suffers from the same problem when it comes to explaining informativeness. This option presupposes much of the causal-historical picture of how names are individuated, but the descriptions that are conceived as characters of names are much less cognitively demanding than descriptions of causal chains. They are metalinguistic descriptions like “[dthat] the bearer of ‘N’”. As it seems clear, this description is a piece of knowledge that everyone acquires upon learning a new name, so they can more plausibly function as cognitively accessible characters. Note that, in option (2), names are specific, i.e., they are individuated in terms of the baptisms which introduced them. As explained earlier, there is no single name “Ludovic” with a single meaning, but many different names spelled “Ludovic”, each with one single meaning. We can express this fact by subscripting the names: “Ludovic₁”, “Ludovic₂”, “Ludovic₃”, etc. So, for example, the character of Ludovic₁ would be “[dthat] the bearer of ‘Ludovic₁’”, which would be different from “the bearer of ‘Ludovic₂’”, and so on. This guarantees that the description picks out the right individual, because the causal chain determines which specific name is loaded into the character. This explains why it can be informative to be told that “Ludovic₁ is Ludovic₂”: their characters are given by different descriptions, since different names occur in them. However, Paderewski cases are unaccountable by this proposal. Since the causal chain pre-semantically individuates one single name “Paderewski”, the same characters would occur in both sides of “Paderewski is Paderewski”. This sentence should be trivial, but it is not.

Let us now look at option (3). As I mentioned, this option treats names as indexicals. This is already very suspicious. As Kaplan says, “those who suggest that proper names are merely one species of indexical depreciate the power and the mystery of the causal chain theory” (Kaplan, 1989a, p. 563). However, let us give it a shot. In this view, a
generic name like “Ludovic” will have a single, context-sensitive character – “[dthat] the bearer of ‘Ludovic’” – which will refer to whoever is called “Ludovic” in the context of its use. In other words, this option attributes to the generic name “Ludovic” one single metalinguistic and context-sensitive meaning, which determines the content in a context. A generic name, then, functions exactly like pure indexicals such as “here” and “I”. Needless to say, this option is not very persuasive in its own right. Generic names do not appear to have meanings by themselves without being associated to a determined object; they are precisely that: generic. Moreover, if there is one single character for every occurrence of “Ludovic”, we will run into the same problems that gave rise to Braun’s revision of Kaplan’s theory. For instance, if we have an utterance of “Ludovic is taller than Ludovic”, we will have to say that the context shifts midsentence, or that the characters of each occurrence of “Ludovic” are different, so that different objects are determined by them; otherwise, this sentence will always come out false. In any case, either the character is the same and the context shifts, and hence there should not be any difference in cognitive value between the two occurrences of “Ludovic”, or their characters are different, and thus the character would not be able to explain the triviality of sentences like “Ludovic is Ludovic”. If this is right, then we can also rule out option (3) as an adequate explanation of cognitive value.

This leaves with option (4). In this view, names function just like indexicals, but their characters vary from speaker to speaker or even for the same speaker in different occasions. Each speaker, then, attributes her own character to a given name at a given context, and this character determines the content. In a sense, this view is very similar to Fregean and Russellian descriptivism: each speaker associates one definite description with a name in a context of use, and this description determines the referent. Just like traditional descriptivism, in option (4) there is no single privileged description community-wise or even speaker-wise; as long as the referent remains the same, there is (apparently) no problem. The main difference is that, contrary to traditional descriptivism, these descriptions are not encoded in the propositional content of sentences containing names; they are turned into directly referential terms and remain only at the level of character. So, for instance, if I say “Socrates is a great person” and associate the description “the greatest soccer player from Ribeirão Preto” with “Socrates”, its character will be something like “[dthat] the greatest soccer player from Ribeirão Preto” for me. The content is just Socrates
himself. Similarly, my audience may associate different descriptions, and thus different characters, with the same name. In short, this view claims that characters vary contextually, given what sort of information the speaker has about the referent and what sort of information is relevant in the context of communication. This is why “Paderewski is Paderewski” can be informative: I associate different descriptions with each occurrence of the name, and thus the character of this sentence could be something like “[dthat] the pianist called ‘Paderewski’ is [dthat] the statesman called ‘Paderewski’”. The propositional content is just a self-identity, but the characters of the two occurrences of the name “Paderewski” are different, so it is an informative self-identity. And this also explains why the same sentence can be trivial: if I associate the same character to both occurrences of “Paderewski”, then “Paderewski is Paderewski” will turn out uninformative to me.

In sum, then, this view claims that the characters of names are determined only in the speaker’s idiolect by the information she has, not by general rules of language. Moreover, proper names turn out to be mere place-holders for definite descriptions, for they do not have constant meanings at the level of character, much like the dthat operator itself. Because this view is so similar to Fregeanism and Russellianism about proper names, they solve Frege’s Puzzle more or less like in the same manner. Additionally, this view manages to avoid Kripke’s modal arguments, because the description associated to a name is not expressed at the level of content. These descriptions are also not merely rigidified: they are turned into genuine directly referential expressions by something like the dthat operator. Finally, contrary to Kaplan’s theory, characters of names will no longer be constant functions, but merely functions from contexts to contents just like any other indexical.

However, despite avoiding Kripke’s modal argument, option (4) does not avoid his semantic argument. If it is my job to associate a description, and thus a character, to a name in an occasion of speech, then my utterance will determine the wrong object if I associate the wrong information with the name. Think of the name “Einstein”. If I associate the description “the father of the atomic bomb” with it, then its character will be “[dthat] the father of the atomic bomb” and would pick out whoever satisfies this description. This character, of course, does not determine Einstein, but Oppenheimer. Therefore, the proposition I express will be a singular proposition about Oppenheimer, and not one about Einstein. This, of course, runs counter well established externalist arguments.
about reference determination for proper names. Moreover, we intuitively say that I have false beliefs about *Einstein*, not that I am thinking and saying things about Oppenheimer. If characters of proper names are determined internalistically, in my idiolect, then it is always possible for my utterances to determine the wrong objects if I have mistaken beliefs about them.

Additionally, this picture makes communication very mysterious. If characters of names are potentially multiplied by all the speakers in a given situation, then how do we explain their grasp of *what is said*? If there is no significant overlap between the information speakers associate to the name uttered in a given occasion, then the route to reference will be so different for each of them that it seems very hard to explain how they arrive at the same content and know that they do so. In fact, this seems rather miraculous. As with indexicals, we intuitively say that is the job of the *speaker* to exploit a single character that must be grasped and interpreted by the audience. In short, reference seems to be a two-place relation between speaker (or expression-in-context) and reference, not an *n*-place relation between every single person in a communication exchange and the referent. Of course, we may associate a huge body of information with a given name of an object, but to claim that we also *express* or *display* this information through an utterance of this name, even if *via* character, is hard to swallow. Worse yet, if I have conflicting individuating information about a given object, I refer successfully to it only if I happen to associate the right description when I use it; otherwise, I will refer to something else. Not even *my own* uses of the name will be consistent in this view.

It seems, then, that all options for explaining the name version of Frege’ Puzzle in terms of character are flawed in some sense. They either fail to account for cognitive value or they seem semantically implausible. Remember also that claiming that names have characters *and* contents just like indexicals is already a *major* deviation from standard referentialism. This lends extra weight to the objections posed above. In the next section, I will briefly examine the purported solution in terms of reflexive content which, as we will see, fares no better than the solution *via* character.

### 3.2 Names and reflexive content

Perry offered two solutions to the name version of Frege’s Puzzle *via* reflexive content, one in Perry (1988) and the other in Perry (1997). Let us start with the former. In
that paper, Perry claims that there is a piece of information that is linked to “Cicero is Tully” that is not linked to “Cicero is Cicero” despite the fact that they express the same proposition: the metalinguistic information that “Cicero” and “Tully” stand for the same thing. This information, then, is not expressed at the level of the official or truth-conditional content, but only at the level of the reflexive content of an utterance of “Cicero is Tully”. In other terms, a speaker who is competent with both “Cicero” and “Tully” may nevertheless acquire new information upon being told that “Cicero is Tully”, and this information is expressed in the reflexive content by containing the very names “Cicero” and “Tully”. This is why “Cicero is Tully” is informative whereas “Cicero is Cicero” is not.

This is all well, except for one thing: contrary to the reflexive content of indexical sentences, this metalinguistic information is not in any sense derived from the meanings of the names in question, i.e., from their rules of usage. Of course, the reflexive proposition “‘Cicero’ and ‘Tully’ refer to the same thing” must be true if “Cicero is Tully” is true as well, but this proposition is not in any sense derived from the linguistic meanings of the expressions involved. It is information about English sentences and names. This is not even a solution in terms of applied linguistic meaning, as was the solution to the indexical version of the puzzle. In short, this is pre-semantic information. Perry, like Kaplan, believes that the only meaning of a name is its referent. Hence, if he wants to explain cognitive value in terms of meaning, he cannot appeal to this kind of metalinguistic information to do so. This information is simply outside the realm of semantics. In fact, this metalinguistic strategy is very similar to Salmon’s (1986) strategy for solving Frege’s Puzzle. As is well known, Salmon does not sympathize with the Fregean criterion of adequacy.

In his (1988) paper, Perry was answering Wettstein’s (1986) challenge. He interpreted Wettstein as claiming that, since the propositional content of “Cicero is Cicero” and “Cicero is Tully” is identical, there is no difference whatsoever between their cognitive value. Thus, there is nothing for the direct reference theorist to worry about. Perry’s solution sketched above clearly meets this challenge. After all, there is a significant difference between the cognitive values of “Cicero is Cicero” and “Cicero is “Tully”, and this fact could possibly be explained by Perry’s reflexive content. But that is not what

Wettstein was claiming in that paper. He was claiming that it is not semantics’ business to explain cognitive value, not that there is no phenomena of cognitive value at all. And this challenge is not successfully met by Perry (1988), since his solution is not based on any semantic properties of proper names. Unfortunately, the solution he offered in his (1997) is not able to meet it either. Let us see why.

In Perry (1997), his theory of reflexive content is a little bit different. He claims that, for an utterance $g$ of the sentence “David uses LISP”, its reflexive content is something like

$$R: \text{There is a person } x \text{ and a convention } C \text{ such that}$$

i) $C$ is exploited by $g$;

ii) $C$ permits one to designate $x$ with “David”;

iii) $x$ uses LISP$^{60}$.

A convention $C$ for Perry is the convention that is established in a baptism ceremony. As he puts it, “When a person or thing is assigned a name, a permissive convention is established: that name may be used to designate that person” (Perry, 1997, p. 6). So, when we use names, we are exploiting these conventions, and it is in virtue of them that we refer to the objects we do. Essentially, then, conventions $C$ are things like Kripke’s causal-historical chains, which determine the referent of a given token of a name, and they are part of a given utterance’s reflexive content $R$. Furthermore, $R$ is known by every competent speaker: even if we do not know who the referent of “David” is, we would still grasp $R$ because we know how language works and how names refer.

This is presumably why sentences like “Cicero is Tully” can be informative (Perry does not explicitly talks about these cases in that paper): these names exploit different conventions, $C$ and $C'$, and this fact is expressed by the reflexive content of the sentence “Cicero is Tully”. Conversely, this also explains why “Cicero is Cicero” is trivial: the same convention is being exploited twice, and this information is encoded in its reflexive content. Since all competent speakers grasp reflexive contents effortlessly, this is not something too esoteric. Contents $R$ like the above are easily available for the speaker’s cognition every time an utterance is produced.

However, like Kaplan, Perry believes that these conventions $C$ are not in any sense an aspect of the *meaning* of the names they introduce. They are not like the characters of indexicals that, in a context, determine the referent. The determination of reference is done *before* any semantic evaluation. Conventions $C$, then, are pre-semantic, as I explained in section 3.1. I quote:

> The role of context in resolving the issues of which naming conventions are being exploited is quite different from its role with indexicals. In the case of indexicals, the meaning of a given expression determines that certain specific contextual relationships to the utterance and utterer—who is speaking, or to whom, or when—determine designation. Different facts are relevant for different indexicals, and the meaning of the indexical determines which. *Names don’t work like this.* The difference between “David” and “Harold” is not that they are tied, *by their meanings*, to different relationships to the utterance or utterer. The role of context is simply to help us narrow down the possibilities for the permissive conventions that are being exploited (Perry, 1997, p. 7, italics mine).

As we can see, Perry believes that conventions $C$ are not in any sense encoded at any level of the meaning of a proper name. They are external relations that fix its reference. Figuring out which relation is being exploited and thus which thing is being referred to is a matter of contextual ingenuity for sure. But this is not something that is determined by meanings, or even by applied meanings, as happens with indexical expressions. Discovering which name and which permissive convention is at play is like figuring out which sense of the word “bank” is intended in a given utterance of “Mary went to the bank”: there are no rules built into the meanings of “bank” that tell us which meaning is being intended in this situation. If this is correct, then Perry’s (1997) proposal may explain cognitive value all right, but not on semantic grounds. Wettstein’s challenge remains unscathed.
4. Conclusion

It seems, then, that neither the character nor reflexive content are able to solve Frege’s Puzzle in any of its versions. If this is really the case, then maybe Wettstein (1986) is right: we have very good reasons to doubt Fregean adequacy criterion for semantics, i.e., the idea that cognitive value is an aspect of meaning. Of course, the arguments in this dissertation are far from being conclusive in this regard. They are not intended to criticize the Fregean approach directly. However, they show that unless the referentialists who accept the Fregean criterion of adequacy come up with completely new theoretical resources, they have little hope for dealing with Frege’s Puzzle on semantic grounds. Character and reflexive content are simply unable to explain all that needs to be explained.

And if Wettstein is right regarding what semantics is about (and I think he is), we must radically reduce our expectations about this discipline. If the proper domain of semantics is only the referential truth conditions of sentences and their literal meanings, as it seems to be the case, then its role in a general theory of communication is much smaller than it was usually supposed, whereas the role of pragmatics is much bigger. In other terms, if the arguments in this dissertation point in the right direction, then it seems that the information we acquire through linguistic meanings alone is much more modest than we believed. The study of meaning, then, does not reveal so much about our thought as it is usually supposed.

Moreover, Wettstein’s most polemic thesis is very friendly to semantic minimalism, i.e., the thesis that the literal meanings of sentences can be determined without any contextual intrusion. In other words, semantic minimalism defends that we should do semantics more formally, at the level of sentences, not at the level of utterances. Semantic minimalism, then, draws a sharp distinction between semantics and pragmatics, and this seems quite compatible with the rejection of the Fregean model of semantics. If the phenomena of cognitive value are accounted for by pragmatics and not by semantics, then the literal meanings of the sentences we utter have a much smaller role in our belief formation, and this is accommodated pretty well by minimalism. In sum, if we reject the Fregean criterion of adequacy, then a minimalist semantics becomes much more interesting, for it is not concerned with explaining anything that is not an aspect of
sentences as types and their literal meanings. But his is of course a matter for future investigations.

All in all, I believe that the deep lesson to be learned here is that if we see meanings as objective and shared things (fitting in what Kaplan calls ‘consumerist semantics’) pertaining to a common realm of semantic facts as direct reference does, then they will never be as fine-grained enough to individuate cognitive value. In other words, if semantic properties are objectively assigned to expressions – whether at the level of sentences or at the level of utterances – they will never be able to accommodate precisely all the possible epistemic profiles of those expressions. There is always the possibility of the same semantic property producing distinct cognitive values for distinct speakers. Cognitive value, then, seems to be an inherently perspectival matter. Unless we have a perspectival – and thus subjectivist – semantics that reflects this fact, such as Fregean semantics, we cannot solve Frege’s Puzzle in terms of meaning. In short, we cannot meet Frege’s criterion of adequacy unless we put meanings back in the head. But we, as referentialists, do not want to do that. Frege’s criterion must go.
5. References


