The many Worlds of meaning: A framework for object reference

Eugene Philalithis

Doctorate

The University of Edinburgh

2019

Name of student:	Eugene Philalithis	UUN	S0831743
University email:	E.Philalithis@sms.ed.ac.uk		
Degree sought:	Doctorate	No. of words in the main text of thesis:	65000
Title of thesis:	The many Worlds of meaning: A framework for object reference.		

Insert the abstract text here - the space will expand as you type.

Our words seem connected with objects – some in our current perceptual experience and some beyond. Words seem to link us to cats, authors, cities and places, present, historical or fantastical. In this work, I try to describe this phenomenon of *reference* in a unified way: one story that can explain the role of words, of things in the world, and of the intermediary mental content often called ‘meaning’ as part of a greater mechanism common to all cases.

This is an old and massive challenge and what I offer here is at most a framework for how that story could be told – one built from pieces already present in the formal, philosophical and (recently) behavioural sciences. I regiment the broader problem into a call for a theory of *contact* between lexical labels and objects in the world, a theory for information *content* attached to labels, and a theory of reference *coordination* between agents or communities.

To construct this framework, I trace historical links between two very different projects: the logico-philosophical ‘classical semantics’ of Frege, Russell and Kripke and the more recent computational-psychological view of conceptual cognition based on generated hypotheses. I argue these two approaches can be seen as continuations of each other: one providing a logical blueprint for what the other already explains and describes. Specifically, I argue the information structures used by causal and/or probabilistic generative models of conceptual cognition could also constitute a theory of ‘meaning’ (in my terms: content) – one linked so closely to a given physical environment that contact and coordination, even as envisioned in classical semantics, will follow from the suitable creation, revision and communication of hypotheses anticipating that environment. I end by presenting two sets of empirical results, on effects from conceptual cognition (categorisation and feature inference) on choices and development of lexical labels in dialogue, relating language use to conceptual coordination.

The lay summary is a brief summary intended to facilitate knowledge transfer and enhance accessibility, therefore the language used should be non-technical and suitable for a general audience.

Name of student:	Eugene Philalithis	UUN	S0831743
University email:	E.Philalithis@sms.ed.ac.uk		
Degree sought:	Doctorate	No. of words in the main text of thesis:	65,000
Title of thesis:	The many Worlds of meaning: A framework for object reference.		

Insert the lay summary text here - the space will expand as you type.

Our words seem connected with objects – some in our current perceptual experience and some beyond. Words seem to link us to cats, authors, cities and places, present, historical or fantastical. In this work, I try to describe this phenomenon of *reference* in a unified way: one story that can explain the role of words, of things in the world, and of the intermediary mental content often called ‘meaning’ as part of a greater mechanism common to all cases.

The focus of my project is on two fronts. I first clarify the questions asked about reference by a particularly important trio of philosophers, into the question of how words connect to things in the world (contact), what sort of information is carried by words when we use them (content), and how we use the same words to mean the same things (contact).

I then attempt to answer two of them using tools already available from empirical science, linking content to a powerful type of recent computational model building full hypothetical objects (which I argue content is most like) and linking the question of how we connect to things in the world with words, to how we can connect to things in the world with our eyes.

I then combine these to consider how content and contact together can bring reference so close to those objects it connects us to that we can replace them with the simulations from that model, for analyses. We can (and I argue do) think of those simulated objects as real.

I lastly show two experiments aiming to connect the third part of the problem, coordination, to some ways that simply talking to each other literally change how we think about things.

I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where states otherwise by reference or acknowledgment, the work presented is entirely my own.

Eugene Philalithis

22/4/2019

Foreword

No doctoral thesis is simple. This one was complex by design, striving to integrate present thinking from scientific psychology with past insight from analytic philosophy, and sketch a new framework for reference, from inert words to living ‘Worlds’ of meaning. And it was complex by circumstance, with crippling stretches of time and effort lost to illness. What is present is the result of compromise.

On the one hand, it only represents a portion of all the work done toward this project, and much that has been sketched here was more developed, but regrettably infeasible to include. On the other hand, it also represents the essential ideas the larger, unseen portion of this work led toward. Although it is disheartening to consider all that got lost along the way, I thus hope that like language and reference, what is immediately present in this compromise might serve as a guide to a greater universe of ideas.

And just like reference, a doctoral thesis is an inherently collective effort. Whether explicitly through our discussions on this project, or implicitly over our ongoing conversations on mind, language, logic, perception and behaviour, none of the following material would have existed in this form without the wisdom, the encouragement, and (ultimately) the faith of the four people who supervised its direction.

I thus extend my deepest gratitude to Holly Branigan, Paul Schweizer, Alistair Isaac, and most of all Andy Clark, without whose unflinching confidence this work would have been sacrificed to necessity. I may not have been the student any of you deserved, but you are each what I aspire to be as a mentor. If on a later day any part of this work can reach its full stature it will be because of what you did right.

For the experimental portion of the project partly reported here, I am grateful to Kiera Davidson and Emma Yapp for their assistance in material preparation, data gathering and variable coding. I remain enduringly grateful to Anita Böröndi for her incisive interventions on my approach to lexical corpora.

I further thank Mark Horne, Nick Treanor, Tom Booth, Alex Doumas and Brian Rabern for giving me a start on serious academic teaching, and for still treating me with fairness, kindness and respect in the toughest of personal circumstances. Though I was not always able to thank you then, I thank you now.

I thank Martin Pickering, Richard Shillcock, Rob McIntosh, Alasdair Richmond, Dave Ward and Inna Kupreeva for their valuable insight and their consistent interest in my work which I hope not to waste. There are countless people with whom I have discussed these ideas (and more besides) over the years of this doctoral project, and five years in Edinburgh beforehand, and I cannot thank them all by name. I am nonetheless incredibly grateful for anyone who has ever had time, comments or criticism for me.

I thank my examiners, Mike Wheeler and Mark Sprevak, for their time and their considered thoughts.

I thank the College of Arts, Humanities and Social Sciences for their many years of generous support, and Tillman Vierkant for fighting my corner, over this funding and elsewhere, since my very first day.

I also greatly thank Alan Bundy for ensuring my next steps on from this project are ever more secure.

Finally, I thank Kiera Davidson (separately) for her time and patience, and my parents for the chance to be here at all, long in the making; for all the cultures they imparted, and the curiosity they instilled. Although only one of you will now read this dissertation, I suspect both of you would have approved.

Contents

Chapter 1: Words, Thoughts, and Things

1 The Problem with Meaning

> “In recent years, […] the existence and importance of this problem of Meaning have been generally admitted, but by some sad chance those who have attempted a solution have too often been forced to relinquish their ambition.” (Ogden & Richards, ‘The Meaning of Meaning’, 1923, p. 1.)

An account of meaning ranks among the most enduring goals in the contemporary study of cognition. From its modern reappearance in Frege the strange relationship of words, thoughts and things which seems to underlie our ability to connect one (words), to the other (things), by means of some sort of cognitive information (thoughts), has lastingly intrigued and puzzled. What do words express? How do words express anything? Is a statement like “ducks quack” about ducks, or about words (‘ducks’ linked to ‘quack’), or about concepts of ducks and quacking? Does a unicorn quack? Should it quack?

As the epigram implies, the analysis of ‘meaning’ has had a long enough modern history but even the most basic question of using some label (as I will call my object of analysis) to refer, and using some sort of system to track what the label could be about has multiple ways to ask it. One way emphasises the anatomy of the mechanism, taken up by philosophers like Brentano (1874) under the umbrella of intentionality (how thought overall is about things) until Frege (1892a) provided a more focused goal of understanding the structure of successful referring expressions and then divided the answer across two equally-weighed components corresponding to the target and the information content of referential expressions. Here I will also study meaning as *reference*, in three intuitive parts: contact (how it gets there), content (the information it has) and coordination (when the same words mean the same).

Starting with this intuitive division, which I will clarify more formally in the next chapter, my aim with this overall project is to understand the ways that an empirical approach, motivated by newer kinds of computational model, can complement the original philosophical analyses produced on the topic of reference by a notably influential sequence of three philosophers, Frege, Russell and Kripke.

I will cover only object labels in particular here: common nouns or proper names or any other type of referring expression whose referent is a concrete object or an object class. And I will view such labels in isolation from greater linguistic structure and any of its effects, as the most basic case of reference. I will also only cover the embodied, situated case of reference as a kind of action by cognitive agents; whether humans or human-like users of language, intending to refer to some object using some label.

In the stereotypically philosophical approach to reference, the standard goal is getting the relation of reference to successfully link a linguistic token (a word) with a material token (a ‘thing’) using some reliable process. The inner workings of this process of getting a thing out of a word are secondary to its reliability, and not usually subject to questions of *cognitive* implementation. What matters is that the linkage connects. Contrasted to this, attempts to address the alternative question of how someone infers the information of a referential expression (vs. how a well-defined referring expression is built) take a different path. Looking at the originator and recipient of the referring expression, this school of thought – driven by evolutionary linguistics and psychology – has tried to understand meaning *in situ*.

Focused on the 'thought' side of the reference equation, earlier semioticist authors like de Saussure (1916) pursued the *how* of referential expressions: the way human agents use them to communicate and the social and cognitive aspects of language, so that "sounds imply movements of speech [while] both, as instruments of thought, imply ideas" (Ogden & Richards, 1923, p. 5, discussing de Saussure). In the psychological tradition this has manifested in the considerable corpus devoted to links between our lexical and conceptual organisation. Starting at a 'mental lexicon' of words in the mind, this work then aims to connect its contents to a complex and ever-evolving ontology of concepts and categories.

The emphasis of this empirical approach is on the property-rich computation of meaning by humans: those processes giving labels what information content they might have in the first place. It should be noted that these processes are inherently cognitive, and equally the nature of this approach. That is to say: they involve some sort of mental processing somewhere to explain how meaning is derived from a label by a human cognitive agent. The distinction between this psychology-driven approach and the earlier philosophical view of the problem of meaning is not one of direction (though philosophy often does begin with objects, while the empirical investigation usually goes the other way around, starting from words) but one of emphasis, in placing the information content first and foremost as its concern. Whereas in philosophy, the specific information attached is secondary to whether (and where) a label connects to a referent: the anatomy of reference comes first. Prima facie, this may imply a productive division of labour between the approaches outlined, a referent-centric story covering the relationship of e.g. sense and referent, while an empirical story covers the complex system of words and thoughts.

In practice this is harder to achieve, often due to differences in the context where these phenomena are studied. Most empirical work on the synchronisation of information content attached to one object has focused on the evolution of stable communication across generations, in the domain of linguistics, and on the 'evolution' of stable communication over the course of goal-oriented dialogue in the domain of psycholinguistics. In both cases, the story is broadly similar: agents start with privately held ways of referring to a particular object, for example a duck (or more appropriately a predator in evolutionary linguistics, or some kind of mundane or abstract item in psycholinguistics). Then the need for them to communicate successfully will subsequently shape the starting points into something common to both. Two agents can have a conversation about referents (e.g. a set of image cards), where the way the two interlocutors describe them (Clark & Wilkes-Gibbs, 1986) and ostensibly think about them (Pickering & Garrod, 2013) becomes progressively more synchronised and structurally minimal over the course of dialogue. The lexical labels used for referents may change based on the audience, or the situation. Labels can change, e.g. from 'shoe' to 'loafer', just by the co-presence of different referents, so that "shoe" is used when the other candidate is a fish, and "loafer" vs. other footwear (Brennan & Clark, 1996). And the changes may last beyond the exchange in which they were made, altering concepts.

It follows naturally that attempts to reconcile 'in situ' empirical investigations of meaning and their structural philosophical counterparts would have to concede an entirely abstract account of meaning. Such an account may still be viable. And if so, it would be completely independent of one involving, at least in any critical capacity, the cognitive or social workings of empirically-tractable agents. This way, what could be solved as a logical puzzle in abstraction may become a more complicated puzzle when examined in situ: in a world of human agents dealing with meaning, using imperfect languages to encode and to express the many entities they cohabit their world with. This is the problem domain.

2 The Meaning of Meaning

In a nutshell, the core strategy of the approach I deploy here against the problem of meaning is to shift the problem of reference almost entirely into the domain of mind and the world of referents instead of the more traditional domain of – formal or natural – language and its structure. Through the following chapters I will consider how traditional questions about (what I call) content and contact can be given convincing and plausible answers, by asking them in the context of how language interacts with mind.

What is useful to consider as a primer is the book-length attempt-cum-mission-statement by Ogden & Richards (1923) to explain the problem of meaning in its entirety, as a *two-step* rather than a one-step function, from words to thoughts, then separately from thoughts to things. This neglected account that its authors call the "contextual theory of reference" is indicative of the views I intend to develop – and will serve to highlight a key way the 'problem of meaning' could indeed be shifted to mind and world.

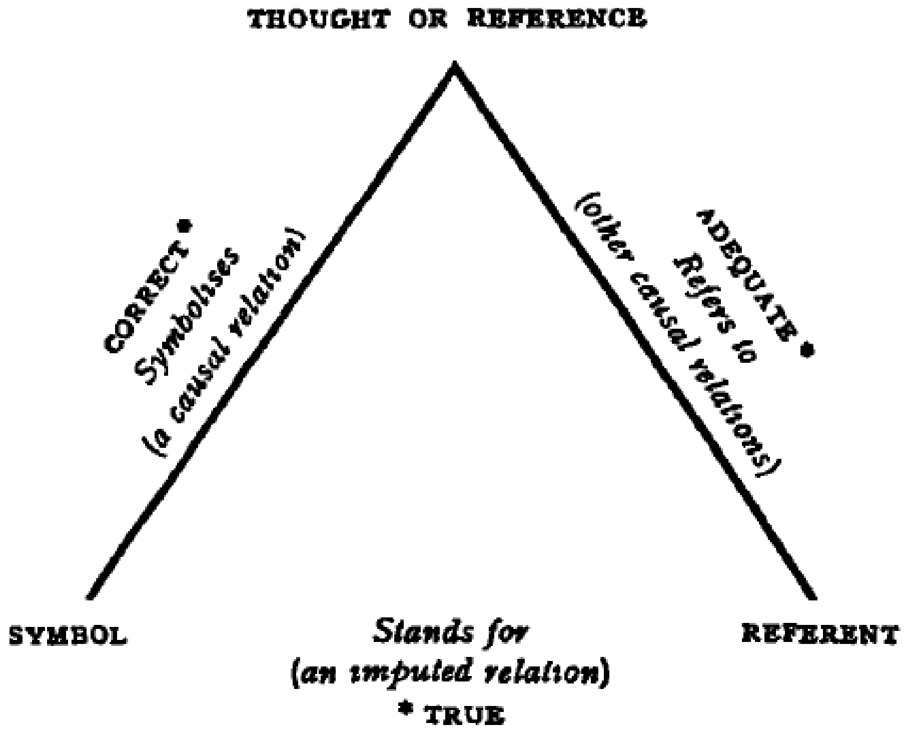


Figure 2. The triangle of reference from Ogden & Richards (1923), p. 11.

At the heart of Ogden & Richards' account of meaning lies the mysterious triangle shown in **figure 2**. The bottom left features the 'symbol', the bottom right features the referent. And nudged between the symbol and referent, the remaining point of the triangle is given as 'thought'. The symbol and referent are separated by an epistemic step so that to get from a symbol to a referent one must stop by thought.

Much hangs on what this stop-over represents. Traditionally a move from symbol that ends in thought may encounter accusations of idealism. Nonetheless this is a triangle with referents still in the (literal) picture, although the direct route across the bottom of the figure is – we are informed by the authors – an 'imputed relation': a virtual, inferred connection. So how does one go from a symbol to a referent through an indirect route, while preserving a genuine connection between them. – in a way that will not "locate Grantchester, Influenza, […] and indeed the whole Universe equally inside [our] head, in such wise that all these objects become conveniently 'mental'" (Ogden & Richards, p. 22). Key to the solution is just what kind of thought this contentious intermediary could be. The cryptic answer given is that this thought represents a 'psychological context' meant to ontologically anticipate the referent.

Translating 'psychological context' and the rest more plainly: what a thought does in this model is set up the brain to expect a particular referent. What it explicitly does not do is replace a referent entirely: the symbol still points toward the external world. Illustrating this is the authors' scenario of striking a match. Striking a match is the causal product of a set of physical movements causing a friction-based chemical reaction between the chemical coating of a match head and some surface, to ignite a match. For Ogden and Richards this is a 'physical context': an arrangement of physical entities carrying out some particular casual interaction. As this match is being struck by a human agent, any movements made and sounds, smell or other accompanying perceptual stimuli cause 'excitation' for their mind. That is, the human striking the match is first-hand experiencing that they are now striking the match. This *experience* of (the sense-data from) a physical context is what they call a psychological context: effectively the echo of a particular physical event or entity in the mind of an agent who experiences it.

Ogden and Richards then associate the two, by exploiting the causal link between a physical context (an arrangement of real entities – the things), and the experience of it (a causally linked arrangement of unspecified mental entities – the thought). Staying with their match example when a match ignites the physical context includes a flame. So the psychological context includes the experience of a flame.

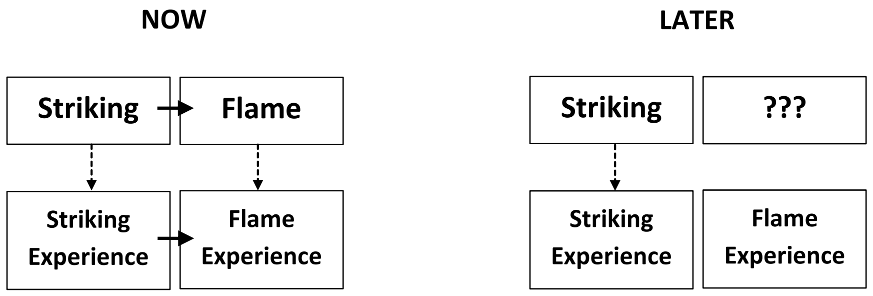


Figure 3: Striking a match according to Ogden & Richards (1923).

Where this account is heading is largely betrayed by the two halves of **Figure 3**. When a match is in the process of being struck, a certain psychological context is created (dotted line vertical arrow), as whoever strikes the match experiences the process. When the match is *actually* struck this causes an adjustment in the physical context (the match ignites and there is now a flame; the horizontal arrow), and the novel part of the physical context (the flame) in turn adjusts the psychological context that a human agent striking a match will experience at that point in time (again: dotted line vertical arrow). But when the same agent starts lighting a match in the future, the learned progression from the strike to the flame in their psychological context will now cause them to expect the flame before it appears.

In much the same way, an agent using a symbol – like the word 'duck' – would come to associate the referent to that symbol internally as part of their psychological context. For example, a duck is often present when someone initially uses the word 'duck'. This way the psychological context of referents and the psychological context of labels (or other signs of referents) become associated independently of the original ('external') causal connection between the physical referents and e.g. letters on a page. The internal relation tracks the external relation, but they do not otherwise interact. Nonetheless, it is that original causal connection in the physical context that lead the psychological echoes of the signs and referents to associate, that creates the reference relation (for this contextual theory of reference).

The authors then expand on this:

> "If we stand in the neighbourhood of a cross road and observe a pedestrian confronted by a notice To Grantchester displayed on a post, we commonly distinguish three important factors in the situation. There is, we are sure, (1) a Sign which (2) refers to a Place and (3) is being interpreted by a person. All situations in which Signs are considered are similar to this. A doctor noting that his patient has a temperature and so forth is said to diagnose his disease as influenza. If we talk like this we do not make it clear that signs are here also involved. Even when we speak of symptoms we often do not think of these as closely related to other groups of signs. But if we say that the doctor interprets the temperature, etc., as a Sign of influenza, we are at any rate on the way to an inquiry as to whether there is anything in common between the manner in which the pedestrian treated the object at the cross road and that in which the doctor treated his thermometer and the flushed countenance." (Ogden & Richards, 1923, p. 21)

What is being glossed in this passage is an account for sign interpretation, where 'sign' can stand for much more than words: such sign situations, as the authors term them, can involve our interpreting a thermometer, or a "flushed countenance", or the familiar rapid movement of a match just as much as "certain letters on a page" or the sign on the road to Grantchester – and in very much the same way. The point thereby implied is that conceptual/linguistic and perceptual processing are commensurable. This is made explicit: "if we realize that in all perception, as distinguished from mere awareness, sign situations are involved, we shall have a new method of approaching problems. [When] we 'perceive' what we name 'a chair,' we are interpreting a certain group of [sense] data and treating them as signs of a referent" (Ogden & Richards, 1923, p. 22). The problem of 'meaning' collapses into perception.

In all, the theory described above is a primitive type of system. Nonetheless, it presents a clear goal to aim for compared to partial solutions available, and its move to collapse all meaning into perception is instructively bold. Obscure though it is, what characterises the attempt by Ogden and Richards is their strong desire to link the components making up the traditional problem and consider how they line up against each other; in particular, how patterns in the world and patterns in the mind can be exchanged.

3 Overview

My approach here will be to sequentially consider first the set of questions from philosophy and then the possible answers from an empirical approach. In the second chapter, I explore the overall style of solution that Frege in particular advanced for the philosophy of reference and the questions stemming from his work, regimented into the separate issues of contact, content and coordination for reference. In the third chapter I consider Russell and descriptivism as a theory of information content, and how a prima facie different theory from empirical psychology may improve it. In the fourth chapter I take on the problem of contact starting with an analysis of perception and then seeking to extend its lessons to reference. In the fifth chapter, I present an overall analytical framework based on the prior work. Then finally I conclude with two sets of experimental results exploring conceptual coordination in dialogue.

Chapter 2: Solving Reference

1 Setting the Stage

This chapter presents reference as a philosophical problem: a puzzle to solve, by way of the classical approach of Frege, Russell and Kripke whose links to certain models of cognition this work explores. I look to Frege for the general blueprint – his overall approach to reference as a problem to be solved generally and logically (vs. heuristically or linguistically), and how empirical input interacts with the logic of his solution to solve the problem of reference coordination, as I roughly stated it at the outset. Having explored the motivation behind Fregean semantics I then more succinctly present Russell and Kripke, and the formulations they each respectively motivate, for the problems of content and contact. Bound up with the problem of 'empty names' these questions set the stage for everything that follows.

Together, Frege, Russell and Kripke have collectively shaped the ongoing philosophical conversation around reference in the 21st century, such that the story from Frege to Russell to Kripke is received as the standard 'creation myth' of the modern status quo for reference – typically with Kripke perceived as the final victor in a century-long series of corrections and replies (e.g. Haack, 1978; Soames, 2007). For present purposes I will suspend judgement on which of these three approaches to reference is best when taken in isolation. With Frege as my starting point, I will instead assess their contributions from a problem-solving perspective: I will explore the major problems that each approach was to solve and the particular theoretical mechanism by which it purports to solve them. In the remaining chapters my goal will be to consider how these mechanisms might all be implemented within the same framework; for which task this chapter sets the stage, introduces key players, and provides the philosophical tools.

2 Solving Coordination

i. First Principles

At the heart of the new ideas Frege brought to the philosophy of language is his contrast of sense and 'reference' – originally found in his "Über Sinn und Bedeutung" (Frege, 1892a) but extending earlier analyses within "Die Grundlagen der Arithmetik" ('A Groundwork for Arithmetic'; Frege, 1884) and "Funktion und Begriff" ('Function and Concept'; Frege, 1891); then itself extended by "Über Begriff und Gegenstand" ('On Concept and Object'; Frege, 1892b) and his later writings (Frege, 1897; 1918). But as the titles for some of this wider literature suggest, Frege came to the topic from a mathematical rather than an empirical or philosophical background. He did not set out to explain language: rather, it became important that he explain language, and reference as a phenomenon, as a side-effect of aiming to explain arithmetic through his logico-mathematical 'Begriffschrift' ('Concept Script'; Frege, 1879).

Begriffschrift was a formal language intended to capture the progression and detail of valid reasoning. Subtitled "a formula language, modelled on that of arithmetic, of pure thought" (Frege, 1879/1967), it was an attempt to make reasoning subject to rigorous mathematical study as a predictable, mechanical process. Starting with well-formed 'thoughts' (I explore this notion more below) that are either true or false, his language tracks the truth or falsehood of their consequences. If thought A is true, denying A is false. If B is false, denying B must be true. These thoughts may be about the natural world, e.g. that opposite magnetic poles attract each other, or about arithmetic, e.g. that 4 + 4 = 3 + 5. As long as they are true or false the language will treat them equally, as a thought *that is true* or a thought *that is false*.

This level of abstraction is important. As long as some thought is well-formed (I will make this exact in §2ii) it is possible to assert that thought is true or false[1] – and this, in turn, is enough for the system to determine the truth (or not) of infinite new thoughts constructed from the original, using operations like negation and conjunction. And this is the hallmark of a system of formal logic. In fact, the system Frege christened 'Begriffschrift' is prototypical of the widely adopted family of (quantified) predicate logics. These are logics of container classes (dubbed "concepts" by Frege) and the potential members of those classes ("objects") and their basic formulae are class membership statements such as 'Kripke is a logician' (true) and 'Kripke is a politician' (false). I revisit applying such logics to reference later. For now, I will dwell on Begriffschrift itself to underscore the type of explanation Frege is aiming for.

Frege aims for a general system of the logico-mathematical phenomena that interest him, from which descriptions of specific phenomena like reasoning or reference automatically follow as a consequence of its rules. Frege does not aim to identify each individual case of valid reasoning. He aims to identify the rules that will produce every case of valid reasoning through a system of mechanical operations on well-defined elements; 'mechanical' so that e.g. applying negation will always flip *any true thought* to false regardless of what the thought expresses. The fuller motivations behind his choice of framework for reasoning are made explicit in the sequel to Begriffschrift, 'A Groundwork for Arithmetic' (Frege, 1884/1953). This work deploys his general formal language for valid reasoning in the specific domain of valid *mathematical* reasoning about numbers – arithmetic. Arithmetic and its formal description are not of immediate concern. However, Frege also ruminates on his general goals and methodology here:

> "I have felt bound to go back rather further into the general logical foundations of our science than perhaps most mathematicians will consider necessary. In the enquiry that follows, I have kept to three fundamental principles: always to separate sharply the psychological from the logical, the subjective from the objective; never to ask for the meaning of a word in isolation, but only in the context of a proposition; never to lose sight of the distinction between concept and object. To those who feel inclined to criticise my definitions as unnatural, I would suggest that the point here is not whether they are natural, but whether they go to the root of the matter." (Frege, 1884/1953, p. 10.)

This is a famous set of fundamental principles, each of which has its own significance, both for Frege and for the core morals I aim to extract from his system. The third of the principles, on distinguishing between concepts and objects in his logical ontology, is important for the logical analysis of 'thought' Frege began with Begriffschrift, then later adapted to use in his account of reference. The second one, on the primacy of propositions over their constituent entities, has attracted considerable attention (e.g. by Resnik, 1980; Evans, 1982; Wright, 1983; Beaney, 1997; and most of all by Dummett, 1971;1993) for its use by Frege in defining natural numbers, in the 'Groundwork for Arithmetic' where it appears. Its precise applicability for his *later* analysis of reference is however a matter of some debate between commentators placing it in the centre of his late philosophy (Dummett, 1993), reformulating it (Evans, 1982), or arguing for its rejection (e.g. Milne, 1986). For present purposes, the single context I discuss the 'context principle' is general enough to rely on a later phrasing (as I will below) such that whether that is *the* context principle, or some weaker reformulation, will not be as important to explore further.

The first principle is the one I take to be most important: that how Frege has characterised thought, or arithmetic, or equally how he characterises reference in work published a few years later, is objective.

[1] The idea that only thoughts with truth values are well-formed thoughts has one important caveat, fiction, that I omit from this discussion for reasons I touch on in section §3i, in the context of what constitutes an empty name.

This is a point Frege returns to repeatedly from his early work on arithmetic through to the end of his career (e.g. Frege, 1918; 1919); and it represents what I take to be the most critical part of his method for present purposes. Derived from general rules, rather than specific heuristics or behaviour, the way Frege analyses reasoning promises complete isolation from the individual psychology of the *reasoner* and, as a result, a guarantee of consistency. No matter who the reasoner is, and what their background knowledge or emotional state, Frege will show how they reason about the same conclusions given the same premises. This method is why Frege can tackle problem of reference coordination so effectively. And his concern to eliminate subjectivity through generality is especially important to how Frege first chose to approach the question of reference with his "Über Sinn und Bedeutung" (Frege, 1892a) – the conventional starting point for any Fregean treatment of the problem of reference, whose own starting point is unsurprisingly (from the above) a question not mainly about language, but about mathematics.

$$(1)\ \ a = a$$
$$(2)\ \ a = b$$

When considering the two statements (1) and (2) above, Frege muses, these must either be statements relating two entities designated by the labels, e.g. the cities Edinburgh and London or even Edinburgh and Edinburgh – or instead statements relating just the two labels '*a*' and '*b*' themselves to each other.

This generates an inconsistency. When the statements relate the labels, (1) is true yet (2) is false given the letter '*a*' is plainly not the same as the letter '*b*'. When the statements relate only what these labels designate, and '*a*' stands for Edinburgh, and '*b*' stands for London, (1) is still true and (2) is still false: Edinburgh (that we called '*a*') is the same city as Edinburgh, but not the same city as London (that we called '*b*'). If '*a*' stands for Edinburgh and '*b*' also stands for Edinburgh however, the labels now both stand for the same thing: Edinburgh features once under the guise of '*a*', then again under the guise of '*b*' despite the two symbols still being different. Based entirely on what they designate, in this case '*a*' and '*b*' mutually standing for Edinburgh, (1) is true and (2) is also true. There is thus *an inconsistency in how the two statements generalise*: if they are taken to relate only the labels, (1) is always true, and (2) is always false. If, on the other hand, they are taken to relate what these labels designate (1) is still always true – whereas (2) can now be true or false depending on what the labels '*a*' and '*b*' designate.

When taken to relate what labels designate, statement (2) is also significant in a way that statement (1) is not. (2) can be either true or false based on the circumstances: as a result, its truth value will encode additional information about these two entities than (1) can offer. It is impossible for something to not be identical with itself. It is quite possible for some entity designated by a label to not be identical to a second entity, designated by a different label. Plainly put: different words could mean different things. And so, the truth or falsehood of statement (2) is not just logically variable in this case, it is also news. (2) being true or false is news. True or false, when an identity statement relates the entities designated by two labels *and* those labels are different as they are in statement (2), the identity statement conveys additional information. It tells us these are two labels for one and the same object. By adding no more than the indiscernibility of identicals (the notion that if $a = b$ then every property of a is a property of b) it says whatever applies to (or is defined for) each of a and b individually must apply to them both. If a is a prime number and b is a Fibonacci number, and $a = b$, then that number is a Fibonacci prime. And crucially, this turn of events requires both parameters to express: both the entities designated and the labels used to designate them must be recorded to grasp this informative use of $a = b$ as a formula, by identifying that case where the *referents* are identical while the *labels* used to express them are not.

To address the divergent logical behaviour of identity statements based on whether they relate entities designated by the symbols, or instead those symbols themselves, and in particular seeing how identity can convey additional information but only as a function of both parameters, Frege bisected reference. One parameter of expressions that refer would be the *mode of presentation*: a record of how a referent was communicated. A second parameter would be the *mode of designation*: a record of only the target entity for the referential expression notwithstanding how it was communicated. The former parameter Frege called the "Sinn" (I will call this 'sense') while the latter he called "Bedeutung" (Frege, 1892a).

To help discuss this material, it will be useful to resolve in advance the question of terminology. This superficial yet surprisingly divisive issue around "Über Sinn und Bedeutung" is the reason I have not translated its title so far and concerns (rather aptly) its labelling in English: the appropriate translation for the mode of designation, "Bedeutung". As Beaney (1997) recounts, all of 'meaning', 'denotation', 'signification' and 'reference' have been suggested to translate 'Bedeutung' and at least one conclave of scholars has been held with the single purpose of agreeing how divergent meanings of 'Bedeutung' as *the act* of referring and also *the thing* referred to, ought to be reconciled using a single term.[2] To be consistent with my presentation throughout, I will continue to use 'referent' for the entities referred to. To forestall ambiguity in later sections, I will also distinguish between two cases for acts of reference themselves (the other meaning of "Bedeutung"): selecting the referent, and attaching information to a referent – a distinction uncertain in Frege as I discuss in §2iv, yet significant in the other two systems. The former case of somehow *selecting a referent* among possible entities I will call 'designation' and the latter case of somehow *attaching information based on some label* I will call 'specification'. Thus given "Edinburgh is the capital of Scotland" the label 'Edinburgh' designates Edinburgh if Edinburgh is its referent. The sentence specifies 'Edinburgh' by attaching the information that it is the capital of Scotland to that label, and (based on the designation) thereby also to its referent the city of Edinburgh. This second distinction is observed by several authors (Donnellan, 1966; Kripke, 1973; Evans, 1982), albeit with different terms and for different reasons, and connects to my proposed distinction between *contact* and information *content* for reference as separable concerns. The ways these elements interact to motivate and to define different systems of reference will be a recurring theme through this chapter.

ii. (Onto)logical Entanglements

Having motivated the distinction between mode of presentation and mode of designation – sense and referent – Frege (1892a) endeavours to clarify their status in his logical universe based on the second of his fundamental principles for Begriffschrift from 'Groundwork for Arithmetic': "never lose sight of distinction between concept and object". What that expresses (in somewhat idiosyncratic terms) is his desire to cleanly divide the entities in his system between predicates – a logical class of container entities – and objects – a logical class of entities the former can validly contain – without overlap. As long as these are distinct, these two logical classes suffice to construct a formal language that records exactly *which* of the objects are contained inside *which* of the predicates. Combined with a system of deduction (cf. Enderton, 2001), this becomes Begriffschrift – or any one of its more familiar modern equivalents, the predicate logics standardly based on a different notation. adapted from Peano (1889) by Whitehead and Russell (1910), like the first-order logics in Enderton (2001) or in Jeffreys (2006).

Each of these possible predicate-object pairings makes for a basic formula in such a logical language, asserting that some object belongs in (the extension of) some predicate. In the modern style of logical notation as opposed to the more opaque notation in Begriffschrift, this basic relationship is expressed by the formula Fa, where F is some predicate in the logical language, and a is a specific well-defined object in whatever domain the language was defined over; whatever set of objects the logic describes. If this is a logic of sheep and a paddock, then a is a particular sheep, and F is the paddock which may

[2] Their consensus was the generic term 'meaning' (Beaney, 1997, p. 36). Unsurprisingly it was soon abandoned.

or may not contain that sheep. *Fa* then means (in English) that the sheep in question is in the paddock. And the list of all pairings of *F* with all the objects standing for each sheep (*Fa*, *Fb*, *Fc*, etc.) together with a record of whether *a* is indeed in *F* or not, *b* is indeed in *F* or not, etc., will exhaust the possible information about those sheep relative to that paddock. Each sheep will have its F-containment status recorded by the logic and there will be nothing more to learn about where all the sheep are being kept.

My intent with this simplistic summa of predicate logic is to illustrate the basic relationship predicate logic is built to express, and its connection to the second principle from 'Groundwork for Arithmetic'. A first-order predicate logic of this sort can describe predicate-object relations and to some extent the requirement Frege places on his system with the second principle is that it simply *be* a predicate logic. That is, to be such that truth assigned to predication statements must categorise every object, as being either *in* or *out*. In the following few paragraphs, I consider some of his reasons driving this choice in a little further detail, to motivate the more exact definitions of the key notions 'object', 'concept' and 'thought' that Frege uses to construct his system of reference. I then move to tackle that system itself.

Frege's intent behind stipulating an ontology of (just) container and contained entities is that together they can express the function-argument relationships found in arithmetic. $f(x) = 2x^3$ is a function that operates on any valid argument in a consistent way. For any valid input it raises it to the third power, then doubles the result. And as Frege (1891) observes in an essay immediately preceding his division of sense and referent, there are key restrictions mathematical function-argument relationships observe. First of all, a function operates on any valid argument in a consistent way, but the output will depend on what particular argument was given. For the argument 1, the value of the above function will be 2. For the argument 2, the value will be 16. For the argument 3 it will be 54, and so on. In each of those cases the procedure followed is consistent and as Frege remarks we can 'recognise the same function' (Frege, 1891/1997, p. 133) underlying the outcomes. It is always raising to the third power, doubling. As a result, there is one part of the structure involved in $f(1) = 1\ 1^3$, $f(2) = 2\ 2^3$, etc. which recurs and another part which is variously 1 or 2 or whatever other valid arithmetical argument we choose for it. It is therefore important to preserve the recurring aspect as a separate concern from the non-recurring when trying to understand how functional expressions operate successfully and clearly represent that.

Secondly, it is possible that the value of a non-monotonic function, like $g(x) = x^2 - 2$, for an argument is the same as the value for another argument. For the argument 2, $g(2) = 2^2 - 2 = 0$. For the argument -2, $g(-2) = (-2) - 2 = 0$. In both cases the outcome was zero but the argument we entered was different. It is thus equally important to preserve the argument as a variable aspect vs. the function computing it and in particular to record whether the same or a different argument was used to compute an outcome, even if that outcome (the value of the function) is the same. This is especially vital for truth-functions, where the value is exactly one of either *True* or *False* for any given formula as input (using this logic) such that tracking the function and its output alone is never enough to fully understand its application: knowing the rules for assigning truth and that some formula is true is not enough to guess the formula.

Thirdly, it is possible that two distinct functions like $g(x) = x^2 - 2$, and $h(x) = x^3 - 2$ will have the same value for the same argument. For the argument 1, $g(1) = 1 - 2 = 1$. For the argument 1, $h(1) =1 - 2 = 1$. It is hence equally important to record when the same or a different function operates on an argument: the familiar issue of knowing whether the same object was being communicated using different labels. The same function can take different arguments while remaining the same function. The values of any one function are not sufficient to determine the exact computation involved. And conversely, identical values for the same argument are not sufficient to know if the exact function being used was the same.

Owing to all the information that needs recording to capture these interactions, Frege (1891) suggests that there should be a representation of a function that identifies *which function* is being implemented, and he calls it a ‘concept’ to underscore its generality. On its own this is clearly not enough to capture all the necessary information. Yet even before this problem of expressive power there is another more purely ontological question about whether any expressions based exclusively on functions could ever stand alone, even before considering the detail such a function-exclusive ontology would be omitting.

Pursuing this question Frege (1891) entertains the notion of picking out purely the recurring aspect of a functional expression toward something like *g( ) = $^{2} + 2$, which is clearly an ill-formed expression. Even beyond any properties of functions that will remain unexplained by omitting arguments entirely, there is no valid use for such an expression in arithmetic: crudely put * $^{2} + 2$ is not a possible number so the function per se is “incomplete, in need of supplementation, or unsaturated” (Frege, 1891/1997). This ‘unsaturated’ aspect of functions motivates the logical picture of function-argument relationships Frege is seeking to construct, by capturing the asymmetry between them. Arguments in arithmetic are numbers (objects) and the values of a function for some arguments are numbers too. The value of f(2) from above is 16, a number. 16 was output by a function given the argument 2, another number, and it can itself be the argument for another function, or even the same function whose output it was. Where 2 and 16 are independent from the function that happened to output one when given the other, and can stand alone, the function (the concept) on its own is simply a blueprint in need of fulfilment This need not mean we must know what an argument is before a function can be analysed. Statements general to *some* or *all* arguments of a function (viz. quantifiers) are one of the headline features of Begriffschrift. It does however *presume* an argument. If no argument can be entered the function collapses – just like * $^{2} + 2$ is not a number. The presumption that properties cannot stand alone will be important for later.

For now, the impact of the function-argument dynamic (expressed using predicates for functions, and objects for their arguments) on the sense-referent dynamic is cashed out immediately in that the sense of any sentence, for Frege, is then simply a saturated function defined in Begriffschrift (Frege, 1892a). Which is to say: the sense of a sentence is a pair of metalinguistic entities in Begriffschrift, rather than the original language, standing in the exact same relation as a mathematical function and its argument. The pair captures whatever ‘meaning’ a sentence had in terms of information attached to its argument. The sense of “Kripke is a logician” is thus a function-argument pair of the function [‘( ) is a logician’] that stands in for the class of logicians, and its argument (‘Kripke’) that stands in for the man, Kripke. This analysis fully extends beyond basic sentences. The sense of “Kripke is either a logician or a fool” is another saturated function: the logical function [‘( ) or ( )’] and its arguments, the sentence ‘Kripke is a logician’, from above, and the sentence ‘Kripke is a fool’ (albeit represented by their truth values: see immediately below). Whether predicate functions with simple objects like *Fa* or logical operators and/or complex objects themselves built of saturated function-argument pairs, like subordinate-clause sentences (Frege, 1892a), all Fregean senses are saturated function-argument pairs. And his term for a particular saturated function-argument pair (in a strictly technical/logical intended usage) is ‘thought’: any given thought is well-formed if and only if it is constituted by a saturated function-argument pair.[3]

The referent of a sentence is then the composite of the (real-world) collection and (real-world) object, whose logical representation is some saturated function-argument pair, represented via its truth value: true if the real-world object mapping to the argument belongs to the real-world collection mapping to the function and false if not. Prima facie Frege has no clear standalone theory of reference for objects,

[3] Strictly this is for “regular” thoughts but I am not concerned with the other (“mock”) type here. See §3i below.

only basic sentences, and I tackle this in the context of his so-called Aristotle footnote in section §2iv. At present, having laid the necessary groundwork, I finally discuss how Frege addresses coordination.

iii. Basing Coordination

Over the two previous sections, I have tried to faithfully render the motivations and methodology that underlie the system Frege first developed for mathematical reasoning, and later adapted for semantics. I strongly emphasised two of the fundamental principles driving his system: the stipulation for object-predicate ontologies to explain function-argument relations, and the requirement for generality. And I considered the depth of commitment Frege displays to a logico-mathematical standard of explanation. My reason for the (still inevitably selective) deep dive into the 'early' Frege of the 1800s was twofold. Firstly, to emphasise the clear mathematical origins of his theory of meaning. That the system he built for reference was not built for natural language reference at all, but rather the relationship between the functions of arithmetic and their arguments; where (as I also noted earlier) certain assumptions can be taken for granted which may not necessarily extend to a different domain. This will be very important for interpreting certain loose ends this system leaves us with as I will attempt to do in the next section. Secondly, although I began with the early and most 'canonical' Frege up to and including "Über Sinn und Bedeutung" (the material typically emphasised by general-purpose textbooks like Soames, 2010), the Fregean semantics I intend to hold up as a template for a system of coordination is instead his last. The theory in question is the most mature – and the most contentious – version of the views in "Über Sinn und Bedeutung", focusing much more heavily not just on the ontology but also the metaphysics of the saturated function-argument pair structure that (per the previous section) Frege labels 'thought'.

Specifically, the (only) Fregean system of reference I will target here is the one reflected in his much later paper "Der Gedanke" ('The Thought'; Frege, 1918) and subsequent contemporary clarifications, like his posthumously-published memorandum to the scientific historian Ludwig Darmstädter (Frege, 1919). And as this is by far the most contentious, or at least the most premise-rich version of the ideas Frege began developing over the material I have already discussed, and continued to develop until his death, it is informative to remember where it all started. By keeping in mind where Frege was coming from as chronicled in the last two sections, his emphasis on a particular ontology, and on generality at all costs, it is arguably easier to understand the stranger places his approach eventually landed him on. Through this section I aim to clarify the role of thought in the Fregean system of meaning, and follow one side of his most contentious argument, to a conclusion which (at least for that side) seems correct. That conclusion supplies the basis for solving the problem of reference coordination (as I will present it via a more precise definition), by securing a logical prerequisite I call commensurable specification. The Fregean endgame for coordination, in the specific context of object labels, is then tackled in §2iii.

Asked to describe what his Begriffschrift is *for* (a non-trivial question, given all the various ways that he described and refined it, explored further e.g. in Sluga, 1980; Diamond, 1986; Heck & May, 2008) Frege is quick to answer with very little preamble, supporting the motivations I have ascribed to him:

> "I started out from mathematics. The most pressing need, it seemed to me, was to provide this science with a better foundation. […] The logical imperfections of language stood in the way of such investigations. I tried to overcome these obstacles with my "Begriffschrift". In this way I was led from mathematics to logic." (Frege, 1919, p. 362.)

He then continues immediately below:

> "What is distinctive about my conception of logic is that *I begin giving pride of place to the word 'true'*, and then immediately go on to introduce a *thought as that to which the question, 'Is it true?' is in principle applicable*. So I do not begin with concepts and then put them together to form a thought or [assign it a truth value]. I come by the parts of a thought by analysing the thought [and its truth conditions]." (Frege, 1919, p. 362, my emphasis.)

This passage, including a (putative[4]) restatement of the ‘context principle’ on the priority of thoughts (as bearers of truth values) over their constituents (the concepts/functions and objects/arguments that make up ‘thoughts’), invites a top-down reading of the fundamental use case for Begriffschrift along the following lines: thoughts are either true or false, and the only kind of thing that can be either true or false. Since thoughts are saturated function-argument pairs, what is present when a thought is true and absent when a thought is false is *the function-argument relationship*. For a basic formula, where the function is a predicate class and the argument is some candidate for inclusion – as with the sheep and paddock – the thought is true if and only if that candidate object is found inside the class – when that sheep is actually in the paddock. Imposing more modern terms for clarity, Fregean thoughts will be true if and only if the object-term satisfies their concept-term, by being a member of its extension. This particular modern terminology is based on Tarski (1933) although Heck and May (forthcoming) argue that Frege dealt with truth in his late work in a way that “has essentially the same purpose, and much the same structure” as the more famous terms and truth theory known from Tarski (1933; 1944).

The only type of thoughts (or formulas) I am concerned with here will be such ‘basic thoughts’ about some object *a* and predicate *F*. Per the above some thought *Fa* is true if and only if *a* is in *F*, else it is false. Thoughts thereby carve the entire world of classes and objects into all the objects in a class and all the objects not in a class, repeated for every class. This method, to classify every object relative to every class, is the fundamental application for Begriffschrift – such that challenges applying it to real number arithmetic effectively ended development of the system (cf. Heck, 2015). What is key for the context of reference is that when combined with truth values, *thoughts classify* their arguments based on the included function. And for Frege, this is the basic case in the application of logic to referential phenomena. Not the objects and classes per se but their classification: putting (or not) the *a*s in the *F*s. In the context of a simple relationship like *Fa* Frege thus mainly analyses the conjecture *that a is in F*. I intend the term ‘conjecture’ as a more intuitive synonym of ‘thought’ rather than some novel entity.[5]

Having delved into ‘thought’ as a logical construct for classification, what is left is to connect it back to the notion of ‘mode of presentation’ (which is a sense; which is a thought) and information content where this discussion began. To find, in other words, the place of Fregean thought in human language. This question is sensitive to a core distinction Frege is observing, between thoughts, that can be given a truth value – e.g. the conjectures that a positive magnetic pole attracts the negative, or $2^2 - 2 = (-2)^2 - 2$, or that Edinburgh is the capital of Scotland – and the manifestations of their conjectures in a human mind, mingled with any “memories of sense impressions” specific to an individual, that he calls *ideas*:

> “Such an idea is often saturated with feeling; the clarity of its separate parts varies and oscillates. The same [thought] is not always connected, even in the same man, with the same idea. The idea is subjective: one man’s idea is not that of another. The result, as a matter of course, a variety of differences in the ideas associated with the same [thought]. A painter, a horseman and a zoologist will probably connect different ideas with the name ‘Bucephalus’.” (Frege, 1892/1997, p. 39.)

This distinction, between thought and *the processing of thought* by a cognitive agent, follows his first fundamental principle “to separate sharply the psychological from the logical, the subjective from the objective” to ensure the definition of the ‘mode of presentation’ is as general as a mathematical proof. What makes the proofs of logic and mathematics general, such that for example one can prove there is no largest prime number, is the assumption of nothing in specific about e.g. *which* prime is considered to be the largest. Taking *some* prime, it is always possible to show it cannot be the largest (cf. Gowers, 2002) without ever showing, or knowing, which specific prime number this is: so this applies to every prime. Equivalently, Frege was not interested in a theory of arithmetic – where he stated this principle

[4] As I said in §2i, my commitment to this being *the* context principle vs. some approximate notion is quite weak.

[5] Intuitive in a scientific or epistemic context, where e.g. $e = mc^2$ is a conjecture that *e* is related to mc^2 this way.

in response to then-prevalent empirical analyses of numbers (cf. Frege, 1894) – or reference at risk of its solutions being limited to a specific context, individual, time, or place. His target was a theory that will maximally generalise like the proof of no largest prime by assuming nothing at all in specific and therefore any mental content *specific to individuals* posed a risk to this goal by virtue of its specificity.

This problem, that Frege identifies in 'Groundwork for Arithmetic' (Frege, 1894), reiterates in "Über Sinn und Bedeutung" (Frege, 1892a – my quote above is from its English translation), then stridently reinforces in 'The Thought' (Frege, 1918) is loosely equivalent to the notion of cognitive penetration in the modern study of visual perception (Newen & Vetter, 2017; Vetter & Newen, 2014). The claim that perception is cognitively penetrated, which I revisit in more familiar contexts in chapters 3 and 4, amounts to the claim that background (e.g. conceptual) information present or accessed at the time of perceptual processing influences the processing outcome. For example, Chalk, Seitz and Series (2010) were able to elicit hallucinations of motion perception, by training participants to expect motion, then eventually perceive motion that was not there. The full scope of cognitive penetration in perception is not wholly unambiguous (cf. Macpherson, 2012) but its existence reflects the sort of worry Frege was expressing: if my percept is 'coloured' by my experience then the percept becomes a joint function of my generic human perceptual processing system and *also* my past experience, which is specific to me.

As a result of this worry, no mental content will be general enough to include in a logical description of reference. And the converse must also be true: nothing general enough to include could be private to an individual cognitive agent. Ideas are individual impressions, including impressions of thoughts, specific to one or another time and place. Whereas thoughts must be the "common store" of mankind (Frege, 1892/1997) if they are to do what Frege meant them to do and underpin a theory of reasoning and reference with mathematical generality based on Begriffschrift. This unusual requirement, for not just the processes but also the *content* of cognition to be subject-independent so it can function as part of a theory of reference, else not be included at all, led Frege to the answer he gives in 'The Thought'. Namely, that thinking (thus by extension reference, whose unit of information content is also thought) must be equivalent to visual object perception, except defined over a nonphysical realm (Frege, 1918). In other words, that thought (and so reference) tracks objects and their relationships in an independent mental space just as perception tracks objects and their relationships in an independent physical space.

This claim, and in particular the requirement that thinking tracks independent thoughts in just the way perceiving tracks independent objects, has prompted much discussion. Whether accepting thoughts as real but only putatively subject-independent entities (Noonan, 1980; Carruthers, 1984), connecting the logical idealisation of thought to 18th-century conventions (Kluge, 1980), naturalising thoughts within language use (Dummett, 1973; 1976), or tethering the information content of every thought to specific physical (vs. general logical) circumstances, or signals (Evans, 1982; McDowell, 1984), it will suffice to say that there are multitudes of opinions on how to avoid the particular conclusion Frege argues for. Setting aside his requirement for a metaphysical substance for 'thought' equivalent to physical objects for perception, which would take me too far afield, and also the requirement for generality for its own sake, I will focus on a pragmatic part of his argument for a subject-independent specification medium.

Specifically, I consider how the Fregean construal of thought licences *information exchange* between thinkers – albeit pending one critical gap in his system to be considered in the next section – and thus invites a more pragmatic (vs. formalistic) reason for the subject-independence of information content, however that subject-independence is realised. Namely how information from multiple sources about the same referent can accumulate and so be contrasted and combined toward a more complete picture.

To gain a clearer view of the argument for thought being objective not only in its logical role (so that a given thought may be the argument of a function in Begriffschrift, and so part of a greater thought[6]) but also its behavioural role, and particularly the intended connection between objective thought and scientific/epistemological use of language, it will help to review a longer passage from Frege (1918):

> "*Is that lime-tree my idea?* By using the expression "that lime-tree" in this question I have really already anticipated the answer, for with this expression *I want to refer to what I see and to what other people can also look at and touch*. There are now two possibilities. If my intention is realized when I refer to something with the expression "that lime-tree" then the thought expressed in the sentence "that lime-tree is my idea" must obviously be negated [i.e. that lime-tree is not just my idea – it is a subject-independent entity]. But if my intention is not realized, *if I only think I see without really seeing*, if on that account the designation "that lime- tree" is empty, then I have gone astray […] If every thought requires [only] a bearer, to the contents of whose consciousness it belongs, then it would be a thought of this bearer only and there would be no science common to many, on which many could work. But [instead] I, perhaps, have my science, namely, a whole of thought whose bearer I am and another person has his. […] No contradiction between the two sciences would then be possible and it would really be idle to dispute about truth, as idle [as for] two people to dispute whether a hundred-mark note were genuine, where each meant the one he himself had in his pocket (Frege, 1918/1956, pp. 300-301; my emphasis.)

Two separate points made within this passage offer a methodological and a pragmatic argument that I take Frege to be giving in addition to the logical one from generality, for subject-independent thought. The first, methodological point is to connect the subject-independence of thought, and the much more plausible subject-independence of perception. If one were told the same lime tree that he perceived in his garden can also be perceived by others, as typically expected of the objects of perception, then all is well. If the lime tree he perceived in his garden is a lime tree only he can apparently see, the worry immediately emerges that he hallucinated it. So long as this contrast is intuitive for perception, Frege argues, the same criterion, the same standard of subject-independence, ought to be applied to thought. If I can prove the same Pythagorean theorem others have proved before, as is typically accepted in the philosophy of mathematics (see e.g. Benacerraf & Putnam, 1983), it is likely the Pythagorean theorem is subject-independent. The Pythagorean theorem can (at the very least) be defined without specifying a particular cognitive agent as its owner, and is therefore prima facie 'immune' to individual variation. Likewise, thought, and thereby the functions and arguments of reference, should have no owner either – they should be 'out there' the same way the lime tree is 'out there' or they could not be constant and therefore could not be trusted. Assuming reference could be trusted, reference must be like perception. Or: since perception can be trusted only when subject-independent the same must be true of reference.

Whether or not that argument is as palatable for non-mathematical entities is not as straightforward as it might appear to someone like Frege who views reference as part of a mathematical system from the outset. The point does remain, however, that either a double standard is applied for perception and for different cognitive activity apparently directed at the outside world like reference, in which case there should be a good theory of why such a double standard applies, or there should be no double standard. I ignore this methodological point for the moment, but revisit it extensively in chapters 4 and 5 below.

The second point Frege makes in this long passage specifically connects to reference coordination. If one cognitive agent, Quine, classifies some object, e.g. (via the thought encoded in) that 'Kripke is a fool' and another agent, Carnap, classifies the same object by saying that 'Kripke is not a fool' (I am taking the negation here for simplicity), then what is it about reference that guarantees that those two classifications of 'Kripke' are classifications for the same object? Aside from utilising the same label (I revisit labels below – though it should be intuitive enough that synonymy does not imply identity), what exactly makes the two conjectures point to the same entity such that the same Kripke, however

[6] As with subordinate clauses, or for intensional contexts like "Frege says ()", which take thoughts as arguments. Frege (1892a) discusses the various ways senses (and so thoughts) can often serve as the argument for functions.

that label is analysed, is or is not a fool, rather than two entities coincidentally called ‘Kripke’ being the classified objects? Similarly, what guarantees the two classifications are linked so one can be the logical negation of the other, rather than being two different and wholly unrelated container classes? For Frege, there seems to be an irrefutable need for a constant ontology underlying both expressions.

This problem is about more than logical resolution, and to explore it further I now revisit and clarify the third of my earlier three intuitive central problems for reference, namely *reference coordination*. The problem of coordination as I expressed it in the last chapter is about how reference can be made consistent between different cognitive agents such that they refer to the same things in the same way. I can now make the requirement more exact using some of the terminology introduced in section §2i:

(C3) COORDINATION

> For a set of labels L, I take reference to be *coordinated* between cognitive agents, if and only if by each label in L the agents i) designate the same entity, and ii) specify it with the same information.

If cognitive agents are coordinated they must somehow refer to the same things in the very same way. There must be no information content attached by one agent to a referent (via its label) another agent will not also attach. And the designated referent, whatever its domain, must be identical for all labels. It is important to emphasise that C3 is a desired end state: the problem is explaining how to get there.

The above is a cleaner definition for the third of my big problems for reference from the first chapter. Satisfying the requirement for all their labels implies two agents have effectively the same semantics. This however is a step further from the situation Frege was describing. The cognitive agents from his example are not coordinated: they plainly disagree on the information content they each attach e.g. to Kripke or to the hundred-mark note in the original passage but they are trying to *become* coordinated. Starting from a state of disagreement over (the classification of) some particular referent, they debate which of them, fake or not fake (re: the hundred-mark-note), fool or not a fool (re: Kripke), is correct. This situation Frege describes is a precursor for coordination. I call this commensurable specification:

(CS) COMMENSURABLE SPECIFICATION

> I call two specifications S_A by cognitive agent A and S_B by cognitive agent B, of referents r and q, *commensurable*, if and only if i) S_A may be substituted for S_B with no change in the information A attaches to r based on its label, and ii) r = q.[7]

Whereas when two cognitive agents are coordinated they refer to the same thing in the very same way and thereby agree on everything, commensurable specification tracks whether any information agents each attach to a referent could be swapped between them, such that A attaches S_B while B attaches S_A, and that both agents may attach information to labels *of the same referent,* no matter how that referent was designated (or whether it was designated by the same process). The intuitive crux of this property is that if specifications are commensurable, they belong to a common ‘information vocabulary’ about some constant underlying ontology shared by the cognitive agents in question. Simply put, the agents may *in principle* classify the same things the same way, which makes it possible for them to disagree: they can make opposite classification decisions about an object. I illustrate this distinction in **figure 1**.

[7] A relative equivalent, without r = q, can also be used but this is insufficient for Frege or for my own purposes. This would mean two cognitive agents can apply the same specifications to the same labels (they have the same language), while offering no guarantee that these labels are linked to the same entities. This is the ‘coherentism’ from Putnam (1981) that I have already briefly considered in the last chapter and I will reject again in chapter 5.

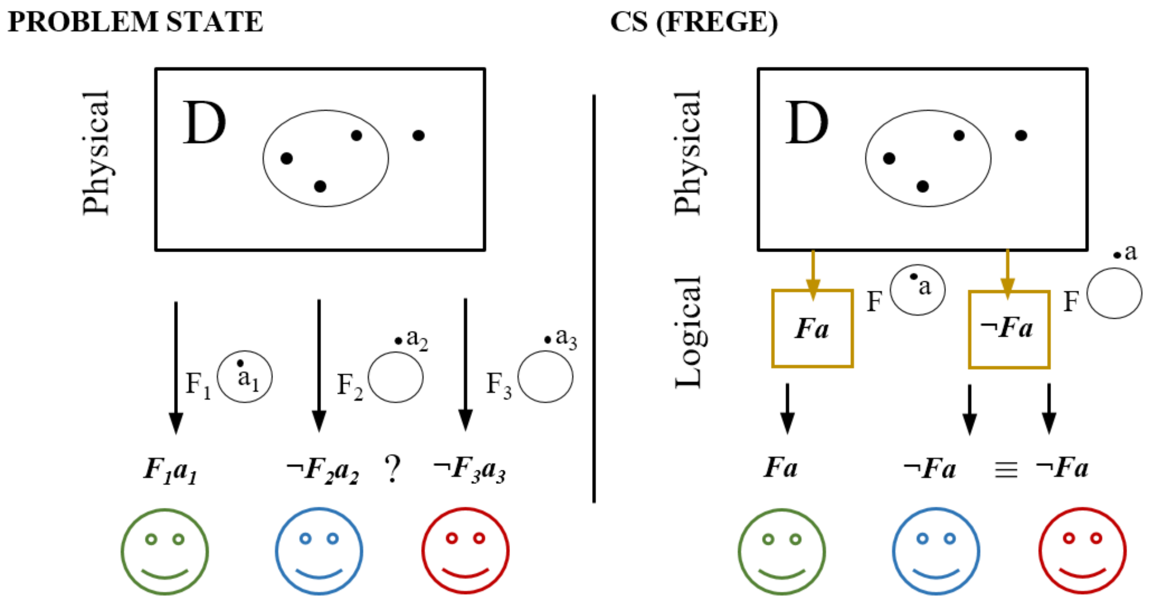


Figure 1. Solipsistic problem state (left) vs. commensurable specification.

On the left is the problem state: the sort of solipsism making reference coordination a problem to start with. Three cognitive agents each attach some sort of information classifying an object in the physical world to a label (not depicted), in the form of a predicate-object pair. But none of the predicate-object specifications are commensurable: there is no determinate logical link between F_1a_1, F_2a_2 and F_3a_3 or any process to determine if a_1, a_2 and a_3 designate the same object or F_1, F_2 and F_3 the same predicate. On the right is a system satisfying CS, specifically as Frege (1918) had envisioned. Between physical objects in the world and the agents attempting to classify them there is a 'third realm' where physical objects and their arrangements are shadowed by logical objects and predicates common for all agents. Through accessing this shared inventory, agents can classify *the same a* as being *in the same F* or not. I set aside for later the link between *a* and the non-italicised 'a' (the original referent being classified) and between *F* and the non-italicised 'F' (the original set of referents forming a class). This is contact.

Although coordinated reference implies commensurable specification, commensurable specification is not enough for coordination. If agents share an information vocabulary and referent ontology they will not necessarily attach the same information to the same entities using that vocabulary. CS is necessary but not sufficient for C3. I illustrate this distinction between systems satisfying CS and C3 in **figure 2**.

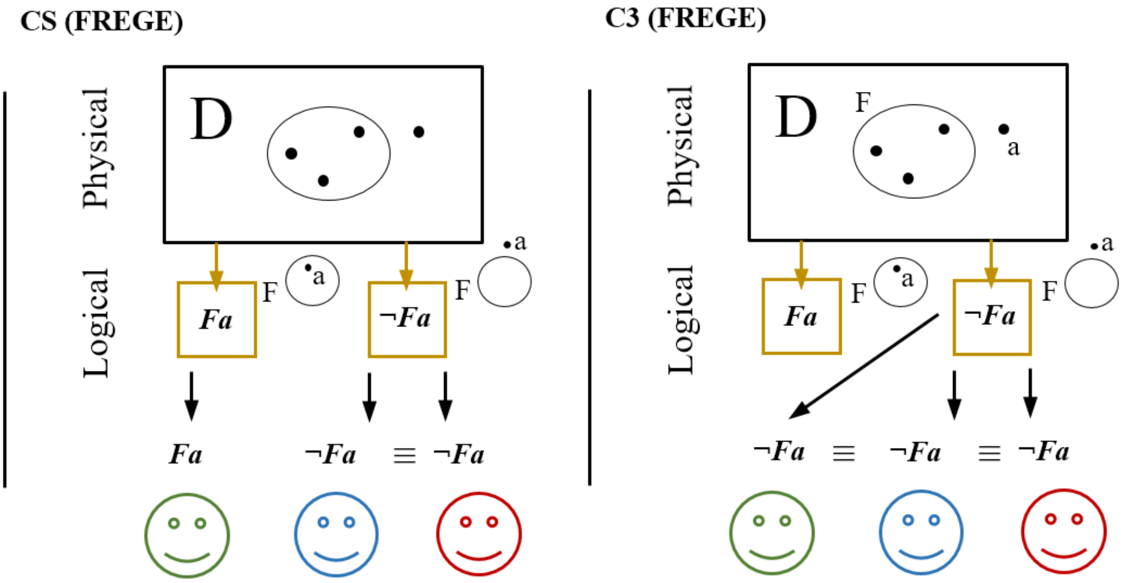


Figure 2. Commensurable specification (left) vs. coordinated reference.

On the left is the CS scenario from figure 1. On the right is a coordinated system of reference where it is not only the case that the agents *can* make the same classification decision for the same referent but also that these agents *do* make the same classification decision for the same referent. This satisfies C3. And I will consider how exactly the move from CS to C3 could be made to happen in the next section.

Coming back to Frege, the two conjectures ‘Kripke is a fool’ and ‘Kripke is not a fool’ thus either *are* commensurable between Quine and Carnap so either conjecture could have been made by either agent with no change in either its designation (what a conjecture is about) or specification (what information content is attached), or instead they *are not* commensurable. Since only commensurable specifications can contradict each other, if instead each conjecture is relative to who said it, so they are talking about ‘Quine’s Kripke’ vs. ‘Carnap’s Kripke’, it does appear “idle to dispute” over who is right. If reference is at all useful in scientific or epistemic contexts – for learning about the world by coordinating beliefs about which objects belong in what classes – it really seems its information content cannot be relative. If Frege is right, it must be linked to the same objects, using a logical lingua franca like Begriffschrift: a perfect mathematical language that, by virtue of its subject-independent generality, offers the bridge. At minimum, some sort of subject-independent information language must be available and some sort of means to ensure information could be attached to (labels of) the same referents must be guaranteed. This stipulation forms the basis for how Frege can solve coordination via his Begriffschrift and it will also form the basis for how I approach the problem later on. Presently I consider the Fregean solution.

iv. Two Mountains

I have extensively delved into the motivations behind the system Frege constructed for reasoning and how that system is generally extended to reference via the division of sense (i.e. thought, a conjecture about a particular function-argument relationship) and referent (i.e. whatever that conjecture is about). And in the last section I emphasised what I take to be his more pragmatic argument that specification must be subject-independent, based on what I called commensurable specification. Without this basis different specifications seem impossible to compare against each other. And so, among other reasons, Frege was led to a subject-independent third realm of thought (sense) to provide specification a basis.

In this section I now attempt to compile a working model of how, on one interpretation, Frege builds on the basis of commensurable specification to solve the problem of reference specification; and how that solution can work for object labels like ‘Kripke’ or ‘cat’ vs. full sentences like “Kripke is a fool”. To get there, I start off by addressing some extra interpretation necessary to reduce Fregean sentence reference into a semantics applicable into my target domain of (only) labels that specify objects. This chiefly concerns the sense of proper names, but I also briefly touch on the role of names qua labels in specification, and how that role in turn affects the relation between the class of proper names (such as ‘Kripke’), that always apply to one referent, and labels that *might* apply to one referent (such as ‘cat’). In the next chapter, I will develop the claim that the two could and should be given the same analysis: but at present I will merely acknowledge the issue of names qua labels in the philosophy of reference. Then I will finally move to how I take Frege to solve coordination – based on a tale of two mountains.

The example set by Frege (1982a) in his ‘Aristotle footnote’ is that the name ‘Aristotle’ is effectively an *alias* for some standard-issue classification sentence such as ‘the man who once taught Alexander the Great’. This has the advantage of straightforwardly casting proper names as an already-explained kind of referential phenomenon rather than taking them to be a separate problem. Every proper name is some sort of ‘trapping’ for the same classificatory conjecture expressed in more elaborate form via some classification sentence – on how some *a* is *F* – of the kind Frege has been analysing throughout. This move is licensed by the consistent way Frege discusses language as the ‘trapping’ (Frege, 1912) for thoughts contained inside, so that different linguistic constructions (or symbols, or signals) can all express the same thought when they all encode the same classificatory conjecture. Though sometimes controversial for more complex cases like time-sensitive language (‘I met Kripke last night’ vs. ‘I met Kripke today’) that have no bearing on my analysis, the policy of treating multiple pieces of language as encoding logically identical information is not exotic. Synonymy is a familiar and common feature of semantics, and Frege effectively formalises this by casting proper names and sentences as different kinds of trapping – or wrapping – for the same logical content which is always expressed as a thought.

What gets in the way of a straightforward analysis of proper names as logical equivalents of saturated function-argument pairs, same as any regular language structure, is that unlike more explicit language structures (e.g. ‘Kripke is a fool’), it is not as obvious *what* the function-argument pair equivalent of a particular proper name (e.g. ‘Aristotle’) should really be. Or as Frege puts it: “opinions as to the sense may differ” between people attaching different information to the name ‘Aristotle'. One person might know him as the man who once taught Alexander the Great, another as the pupil of Plato. And if both people claimed that ‘Aristotle was born in Stagira’ they would each attach information about his birth to a slightly different logical entity. One would be saying that *the man who once taught Alexander the Great* (= Aristotle) was born in Stagira. The other would be saying that *the pupil of Plato* (= Aristotle) was born in Stagira. Although these are both about Aristotle, they are not aliases for the same thought.

Such “variations in sense” are acceptable for Frege though at the same time “they are to be avoided in the theoretical structure of a demonstrative science” (Frege, 1982a, p. 37). On the one hand Frege will accept that proper names, if they are aliases for more explicit descriptions, have senses that could vary based on who attaches them and (crucially) what information might be available to them. On the other hand, he says this situation should be avoided where proper names are used for demonstrative science: if our goal is to classify Aristotle scientifically there should only be a single classification. And that is as far as the ‘Aristotle footnote’ informs us on what Frege had in mind for proper names, though later sources help fill the gaps in what I will call the Fregean *provisional pluralism* for proper name senses.

‘Provisional’ because this is not the ideal language; and ‘pluralism’ because the two specifications of Aristotle from the footnote are both true of the same referent despite containing different information, as opposed to being a correct specification and an incorrect specification, or there being two referents.

Whether the sense of proper names is well-defined by Frege to begin with is an open question and the subject of much debate, particularly as it intersects the debate on the subject-independence of thought. There are broadly diverging opinions, about extending through language (Dummett, 1971), or instead naturalising (Burge, 1979; Evans, 1982; McDowell, 1984) the definition of thought and sense, so that senses for proper names are sufficiently intelligible. With ‘sufficiently intelligible’ I mean describing how senses can be assigned to proper names at all, if not by consulting some true classification of the referent – its logical address in the third realm – ‘magically’ based on its (prima facie arbitrary) label. For linguistic or mathematical structures wearing their logic on their sleeve like ‘x + 2 = 100’ there is an explicit trail of logic taken to *be* the sense by linguistic-turn philosophers such as Dummett (1971). Yet Aristotle has no obvious logical trail. Therefore, the question is how to unearth a purely linguistic trail for Aristotle, or else appeal to some other sort of trail to ‘lock in’ the sense of Aristotle and other proper names e.g. based on a history of learning or of interaction with the referent. And after that step there is the follow-up question of how that process can produce multiple senses for some referent, for example according to the natural language context, or based on one vs. another history of interactions.

There is also a formal question over whether a method exists to identify a logical alias for each proper name that will not itself include another proper name, like Frege saying Aristotle was a pupil of *Plato*, lacking which any attempts to replace all proper names with their logical aliases can regress infinitely. And there is the related question of whether a proper name itself, the lexical label ‘Aristotle’ is part of its information content as Burge (1973) in particular argues for, or is instead eliminated by the logical structure for which it was an arbitrary alias. This question in particular will be important for chapter 3, and I will not pursue it further here except to say that this is an open question; and that if the labels for proper names are themselves part of the information content they attach to their referent, and the same applies to names like ‘cat’ that may (or not) be attached to a single object, then the difference between labels for proper names and for common nouns might be expressed through the size of their extension. Therefore, if all labels are indeed classes, and the information content of ‘Aristotle’ includes *that he is called Aristotle* without circularity, then proper names and common names can fall under one analysis – or so I will argue in the next chapter. For now, I will assume it, and ignore any worries over regress.

It should be clear the above is entering territory Frege did not cover: whether searching language for a trail to the sense of proper names, or searching experience, these complications force a departure from the Aristotle footnote and what Frege has explicitly said about the matter. Much of what is raised over proper name senses also hangs on the other question, of how senses become associated with a specific referent. In other words, how we attach some specification to an arbitrary entity we label ‘Aristotle’ – for which it can then function as a logical shorthand – and not some different entity, or no entity at all. This other question is what I have called the problem of *contact* and I will not discuss it through Frege although many of the same questions will resurface when I finally do consider it here and in chapter 4.

These caveats stated, what I instead return to is coordination: how Frege can make the move from CS to C3 in figure 2, and solve the problem of reference coordination based on his ‘provisional pluralism’ as I have given it from the Aristotle footnote, and pace the numerous complications for proper names. This will not be an attempt at definitive interpretation, and I will consciously sidestep a lot – if not all – the issues I have just raised. My focus will be on the plausible reconstruction of what Frege wanted, or perhaps hoped, to achieve with his system of reference for proper names (thus all names, given my working assumption). More specifically: how his desire for a system of reference accounting for what is true intersects with the epistemic requirement that sense accounts for what is just *thought* to be true.

I pursue this attempted reconstruction from a source familiar to Fregean scholarship (cf. Evans, 1982) telling the tale of the two mountains ‘Afla’ and ‘Ateb’ and the two explorers scaling them for science:

> “Let us suppose an explorer travelling in an unexplored country sees a high snow-capped mountain in the northern horizon. By making enquiries among the natives he learns that its name is ‘Afla’. By sighting it from different points he determines its position as exactly as possible, enters it in a map, and writes in his diary: ‘Afla is at least 5000 metres high.’ Another explorer sees a snow-capped mountain on the southern horizon and learns that it is called ‘Ateb’. He enters it in his map under this name. *Later comparison shows that both explorers saw the same mountain.* Now the content of this proposition ‘Ateb is Afla’ is far from being a mere consequence of the principle of identity, but contains *a valuable piece of geographical knowledge*. […] An object can be determined in different ways and every one of them [corresponds to a name], and these different names have different senses; for it is not self-evident that it is the same object which is being determined in different ways. […] Now if the sense of a name were something subjective […] a common store of thoughts, a common science would be impossible. It would be impossible for something one person said to contradict what another said.” (Frege, 1914/1996, pp. 320-321, my emphasis.)

The plot of this story from a letter Frege wrote to Philip Jourdain is by no means novel to that source. It follows in the tradition of several examples from his earlier work using zoology, astronomy (Frege, 1892a) and mountaineering (e.g. Frege, 1904) to embellish the initial case study he began his analysis of reference with: *a* = *b*. The names of these mountains, ‘Afla’ and ‘Ateb, are anagrams of ‘Alfa’ and ‘Beta’. They dispense outright with mere allusion to mathematical form, like his more famous case of the Morning Star and Evening Star (Frege, 1892a). Mt. Alfa and Mt. Beta explicitly recapitulate what Frege seems to consistently consider the basic case for the analysis of *reference* as part of his broader theory of thought. Namely, the coordination of classifying theories about objects of the natural world.

Be they mountains or planets or animals (cf. the quote about the horse Bucephalus in §2ii), the thread running through these similar examples is his appeal to an underlying motive of unifying information: not just that *a* = *b* is a puzzle, but that this is *the relevant puzzle* to understand what reference ought to do for us, and what a theory of reference most ought to cover, by describing what happens when there are two different pieces of information attached to two different labels both pointing at the same thing. And what (per the story) happens in that case is that we want to compare that information. We want to unify prior knowledge about the classification of *a*, with prior knowledge about the classification of *b*, and thereby come to extend the classification of the object they both refer to. In other words, we want to transfer our discoveries, from one object of thought, and maybe from one thinker, to another object of thought, and possibly another thinker. What Frege is motivating is a logical theory of reference for the use of a common science (this much is spelled out clearly), and transfer of information is what the science in question (based on all the passages I have cited) appears to run on. This is another guise for what I have earlier called the pragmatic side of Frege, to the extent later examples like the debate over the hundred-mark note are intended to cohere with his theory of reference; and there is no cause to see his remarks in ‘The Thought’ as motivating anything but that same consistently intended ‘big picture’.

The intended big picture for Fregean reference based on all of the above might be something like this: one scientist, Dr Alfa, and another scientist, Dr Beta, each discover something novel. (Of course, I am re-treading the standard Fregean story here.) Dr Alfa discovers that chemical *a* has the property *F* and Dr Beta discovers that chemical *b* has the property *F* as well. Dr Gamma – a recent transfer – suspects that a third chemical *c* does not have the property *F*, e.g. because of some other property *G* that *c* does have according to his own research. Dr Alfa thinks *Fa*. Dr Beta thinks *Fb*. Dr Gamma thinks that ¬*Fc*, because he also thinks *Gc*. It is important to note the subtle but important shift from what is true to the belief that something is true which is essential to any one of these examples.[8] Having these beliefs the three scientists meet at their annual conference and this is where the Fregean big picture I have argued for (or at least have inferred) is in play. What kicks things off is the shocking discovery that $a = c = b$.

If we can assume CS about the scientists such that *'a', 'b', 'c', 'F'* and *'G'* are all commensurable and we adopt a pragmatic point of view, the next question is how they can use that property to better grasp the world from the starting point of their beliefs at the point of discovery. Until that point of discovery the scientists collectively employed[9] the sort of non-ideal language tolerated in the 'Aristotle footnote' where a single object is variously specified: they were in that state I called provisional pluralism. And after that, like the two mountaineers from the story the scientists presumably compare notes. Because CS is true they are able to compare their classifications directly. Alfa and Beta can confirm with each other that *Fa* and (crucially) they can compare their belief that *Fa* with Gamma, who believes that *Fa* is not true – and at any rate is absolutely certain *Ga* is the case. Note again the nuance that has slipped in here, where now the beliefs have degrees. I introduce it for a reason that will become apparent later. Doing what they each could, the trio made certain discoveries and certain hypotheses and now attempt to compile them all into just one model. If Alfa and Beta are right this model contains *Fa* – and what I have been describing via this story is that exact transition effected from the left to the right of figure 2.

Classifications were compared until a consensus won, then it was spread across the whole population. And that is the endgame I extract from Frege about reference coordination: that given CS or a similar guarantee of a common vocabulary and a scientific context, *communication begets coordination*. And if it seems like this endgame is arbitrarily imposed on Frege there is one more thing to consider. If all the scientists having compiled their discoveries and expertise now agree that *Fa*, and if they keep at it until after a very long time they classified *a* for every property, then every referent for every property, then they will also have transitioned, from the tolerated provisional pluralism of the Aristotle footnote (their starting state), into the perfect language Frege (1892a) wrote "Über Sinn und Bedeutung" about. Coordination is embedded in a process whose end state is the perfect logical language, not such that it entails that perfect language exactly and parsimoniously classifying the universe but such that getting coordinated gives science an opportunity to make progress toward it: that is the endgame of reference.

What I have thus inferred about this 'pragmatist Frege' is the missing link, between what he describes as the desired end point in "Über Sinn und Bedeutung" and what he later motivates as the desired start point in 'The Thought'. I have not inferred the start or the end point, both of which are vividly spelled out across these and other sources. There is doubtless more to achieving a 'perfected' science than the capacity to coordinate beliefs about classes. Still, Frege (explicitly in 'The Thought') felt this capacity is critical. This is what I imply with the slogan 'communication begets coordination'. That given a) an assumption equivalent to CS, and also b) a process whereby beliefs are compiled to weed out the false and add up the true like the ideal Fregean natural science, *then* joint classification begets coordination. And where for Frege CS is underwritten by subject-independence for those logical objects comprising the particular saturated function-argument configuration he calls 'thought', the question remains open whether this is the only way to underwrite CS and even if it is, whether such an approach can succeed. As a result, my take-home message for coordination in particular is that, *here is a case where it works* by assuming a particular sort of logical ontology (and to an extent a metaphysics) for object reference. As to how or where I think that case is plausible, or feasible, this will be a discussion to return to later.

[8] What I mean by this is it is unlikely that Frege wanted to describe reference in a science where classification is certain. If so it would be pointless to discuss disagreement over classification such as for the hundred-mark note.

[9] CS is what implies this is one language, and I have assumed CS as Frege does.

3 Solving the Rest

i. Running on Empty

So far in this chapter I have taken aim at how Frege frames reference as a logico-mathematical puzzle and then attempts to solve that puzzle, with a consistent emphasis on generality and coordination. The Fregean project to address reference coordination based on a subject-independent logical ontology of information content has been the focus of my attention, and what I aimed to motivate and reconstruct. I now shift my focus to the other two parts of what I identified as the tripartite problem of reference – the problem of *contact* and (completeness of) information *content* – and two authors who emphasised each of them in their contributions to semantics as much as Frege had emphasised coordination in his. Bertrand Russell on content and Saul Kripke on contact round off my tour of 'classical' philosophical semantics; that influential body of competing philosophical systems of reference whose connection to a particular class of modern theories of cognition and perception I have set out to explore in this work. Taken together these three authors also round off an inventory of philosophical questions and answers about reference that provide the exact basis, and the measure of success, for my suggested framework. In other words, they mark the span of my first 'world of meaning' (in the lowercase-w sense): not just as an influential body of theory but as a philosophical *premise set* forming a portion of a larger puzzle.

Following Frege and coordination, the task of more exactly setting out the stipulations for content and contact from the viewpoints of Russell and Kripke is comparatively easier. These authors present their concerns for the problems I will be regimenting as 'content' and 'contact' in a briefer and vastly more self-contained format than Frege. And any connections between their views, if taken as stipulations or premises, and empirical work answering or complementing them are also much more easily identified. Whereas for Frege any links between coordination in his work and empirical science are more diffuse, the questions raised by Russell and Kripke are (I will argue) intimately linked with particular research programmes[10] from modern psychological science. And these links are easier to show alongside those empirical theories that (again, I will argue) appear to either modify, or to outright meet their demands.

For the latter reason in particular, I will thus divide my discussion of Russell and Kripke between this and the following two chapters. Over the rest of this chapter, I will only consider the *problems* raised by Russell and Kripke in response to Frege that I take to comprise the problem of content and contact. As with coordination, my aim will be to extract each problem (and any implied conditions for solving it, like CS) from whatever philosophical context it was raised in. Since Frege is the common target of both these problems, the only context needed is the Fregean system I have already presented at length. It is therefore possible to sufficiently motivate and define the problems of content and coordination in (just) the shadow of Frege without great consideration of their own internal problems and interactions. Over the next two chapters, I will then delve deeper into each problem, and in particular each *solution* given by Russell and Kripke, with their internal weaknesses and connections to psychological science.

For now, I discuss just the problems: what is missing so that a coordination-centric Fregean semantics is not the final word on reference. And one special object of analysis in particular, missing most of all from Fregean semantics, is the *empty name*. This odd object stems from questions of contact – i.e. the 'getting there' of reference – but touches on every part of the fuller problem. And it can serve as a test for whether a system of reference passes muster, as a formal theory, by analysing it successfully. The empty name is specifically the weakness of Fregean semantics that both Russell and Kripke exploited, to launch their competing alternatives (albeit in different ways). And it is the springboard, or test case,

[10] In the Lakatosian sense of an overlying goal or narrative pursued by a research community; cf. Lakatos, 1976.

for many more systems of reference past the ‘creation myth’; e.g. by Meinong (1905), Quine (1948), Zalta (1983), Sainsbury (2005) or Recanati (2012). Empty names are thus a diachronically consistent plot device for theories of reference, directing discussion toward or away from some types of analysis.

To better inform my discussion of Frege, Russell and Kripke, and also systems of reference in general (including the framework I will be suggesting), in this section I will thus attempt to taxonomise all the various instances of ‘empty names’ in a generally usable way (i.e. one not bound to a specific formal system), with the aim of making it clear where each ‘species’ arises, and so where precisely a system of reference ought to worry about tackling it. My taxonomy largely follows the intuitively-motivated style from Kripke (2013) stemming from how different ‘empty names’ are used, though I will present a slightly more elaborated breakdown and will not necessarily accept the same borders between cases. That said, the prima facie idea of an ‘empty name’ can already be explained with the terms I have set out so far: an empty name is a label without designation. Which is to say, an empty name is a name – in my terms a label as I do not distinguish between proper names and common nouns (cf. §2iv above) – that *has no referent*. Most of the trouble with empty names consists in making this condition clearer.

At first reading ‘the present King of France’, ‘square circle’ and ‘Odysseus’ would all be traditionally considered empty, because in each case (at least at first reading) there does not appear to be a referent for them. Cases like ‘unicorn’ and ‘dodo’ might also fit this bill, as there are none of those either. The same might also be said of ‘phlogiston’, where a referent was anticipated but was never there after all (cf. Godfrey-Smith, 2003). And lastly, perhaps a description for ‘the man holding a martini’ in a busy room is an empty name when, instead, there is only a man holding a glass of water (Donnellan, 1966). What I intend with these examples is to showcase the many ways in which there may not be a referent for a label – be that label wrapped as a name or more explicit as a description (again cf. §2iv above) – which are not necessarily tokens of the same type of phenomenon, even if these are all ‘empty names’ in the sense that there is nothing (currently) in the external world of physical objects that they refer to.

Although the conditions for some of these being empty or not are intimately tied up with descriptivist (vs. Fregean) reference, which I will preview in the next section and then tackle in earnest throughout chapter 3, there is room for a more general-purpose classification of typical ways names can be empty even in advance of considering descriptivism. I will follow Kripke (2013) here in distinguishing types of empty names by usage but I will also include a constrained sort of counterfactual usage – a ‘degree of emptiness’. Based on both these considerations I offer a usage-based classification of empty names, and in that process introduce an important element in both Fregean and Kripkean semantics: intention.

First and foremost, there is the class of names intended to be empty. These are the fictional names like ‘Odysseus’ (which is an example from Frege, 1892a), or ‘pegasus’, ‘unicorn’ and so on, as well as the fictional contexts of usage for otherwise nonempty names like a story incorporating a living, breathing individual as a character[11] in an otherwise (intentionally) fictional narrative. This fictional-usage class of empty names has a wide and varied literature; so, as my focus is squarely on reference in its natural (or naturalistic) usage, intentionally empty reference will only feature briefly, to motivate future work. Setting aside the fictional case to focus on labels intended to refer while somehow not referring (in my terms: intended to designate a referent, but not designating a referent at all), there is then a question of how good a job they could have done picking out a referent. Again, specifying this notion would need at least a definition of contact (which I provide later in this chapter, cf. §3iii) but there is an intuitively greater problem with e.g. ‘square circle’ when used as a label, than there is with ‘phlogiston’. There is perhaps a lesser problem with ‘dodo’ or even the otherwise-notorious ‘present King of France’, where

[11] Whether these may be non-empty while being fictional as Kripke (2013) contends will not be of concern here.

the name is empty now, yet would have picked out a living, breathing thing as intended 250 years ago. And there is perhaps no problem at all with ‘the man holding a martini’ even if his drink is not correct – as long as there is no other man holding a glass we could misidentify him as, in that very busy room.

The above distinctions suggest an effect of a particular sort of counterfactual context in resolving, and thus classifying empty names. Kripke (2013) argues the man holding the martini is not an empty name at all as we are still able to pick out the intended man in the room using that label; he also claims there is no distinction between fiction and unintentionally empty names, like the others, where the name has been discovered to be empty. For Kripke, there is a distinction between names *understood to be empty* (whether in advance or post hoc like ‘phlogiston’) and names *understood to be non-empty* by their use despite any apparent misalignment of explicit information content as with ‘the man holding a martini’. The slogan here could easily have been, names either work or they don’t. And when they fail to help a human pick out a referent that (and only that) is what makes them empty. Setting aside how labels not intended to pick out a referent, in advance like ‘Odysseus’ or in retrospect like ‘phlogiston’, could still have an indirect usage, the core distinction Kripke is observing really falls between the names that are successfully used for the job of referring, and names that are not. And what is a successful and helpful usage for picking out the referent in one context by whatever means, could be unsuccessful in another.

Whether some label is an empty name hence resolves into a context-sensitive question of whether that label was (not) successfully *used to designate* a referent. Taking a cue from this idea by Kripke (2013) I classify empty names along three ‘species’ based on those contexts where they are successfully used to designate a referent for some given natural context – where by ‘natural’ I mean a context that either can be found in the current, actual natural world, like a room or (more broadly) an environment or one that could be found in a past iteration of the same. *Mistaken names* are what I will call labels that have no successful use in a natural context – whether explicitly like ‘square circle’[12] or following empirical investigation like ‘phlogiston’. These are labels for which there is no good use at all. There never was a referent. *Vacant names* are labels that have no successful use in this context, but do have one natural (past or present) context where they successfully refer. ‘Dodo’ is a vacant name now as there are none around yet in 1750 it could be used to designate a particular living, breathing bird. *Inconsistent names* are finally labels that have a successful use in the given context, but that use is inconsistent with their explicit (and/or implicit; I revisit this in the next chapter) information content – like ‘the man holding a martini’ where the label is successfully used, but inconsistent with the explicit information attached.

And underwriting this taxonomy is the intent to successfully refer. As Kripke (2013) rightly points out when (I think wrongly) conflating the analyses of mistaken names and fictional names, there is a clear and important distinction, between a label used under the assumption that it does designate something, and a label that is completely ‘given up’ as falsehood or fiction, for how that label should be analysed. It is certainly not the case that fiction consists of ‘failed facts’. It is also not correct that an abandoned label should count alongside labels earnestly meant to designate referents. And for exactly that reason, I would argue that the epistemic transition, from *intending to refer with a mistaken name*, to *giving up a label because it is mistaken*, is something worth preserving as distinct from the purely fictional case.

Armed with this taxonomy of empty names, the remainder of the tools and notions I require are much more easily presented. And easiest of all given the above is the main problem with Fregean semantics: that in aiming to analyse only the endpoint of the ideal logical language arrived at via communication, Frege (1892a) conflates the *intention to refer* with successfully referring, in natural/scientific contexts.

[12] There is a plausible ironic/indirect context of successful usage for ‘square circle’, designating a confusingly complex shape. But this falls under the category of intentionally empty names feeding e.g. Gricean implicature.

That is: the intention to classify *a* as being in *F* by using the expression ‘*Fa*’ (as opposed to a fictional usage which he does cover) entails that *a* designates a referent for Fregean semantics. The conjectural element of ‘*Fa*’ is exclusively about whether that *a* is correctly classified for *F*, under the assumption that *a* is an entity previously (or subsequently) classified relative to other predicates. There is no case described in Fregean semantics where intention to refer does not by itself presuppose a valid referent.

ii. Lost and Found

To make the above more precise: there is no *consistent* case described in Fregean semantics where the use of a conjecture ‘*Fa’* to classify an object via a truth value does not also presuppose *a* is nonempty. In other words, there are only ever two options in play when *Fa* is concerned, as depicted by **figure 3**.

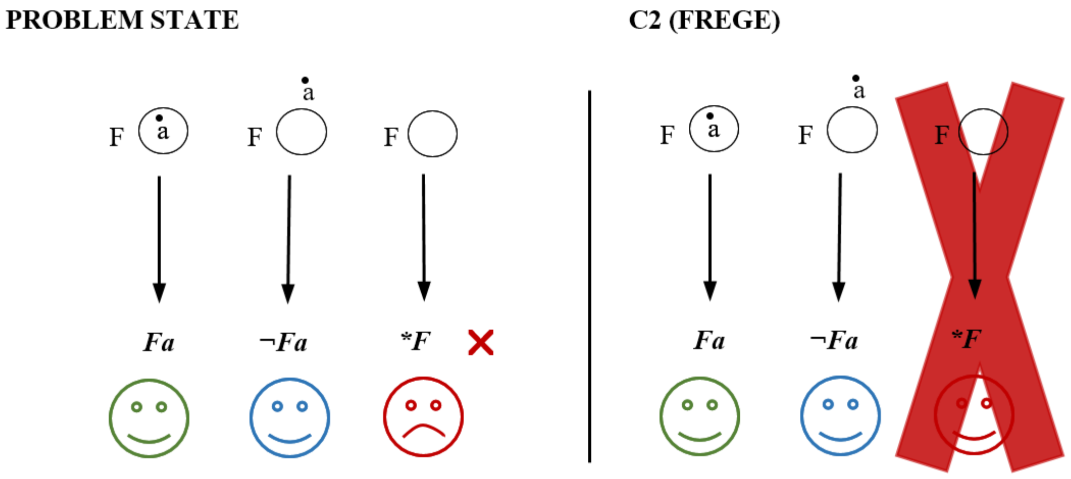


Figure 3. The empty name problem (left) and the axiomatic exclusion solution.

Frege (1884) does allow for the possibility of empty names but the formal structure he assigns to them makes all classification statements involving empty names axiomatically false. This renders *Fa* and its logical opposite ¬*Fa* well-defined in Begriffschrift, when *a* is empty. It also makes both of them false. The result, which Russell (1905) is eager to point out in his deconstruction of Fregean semantics when considering empty names in the picture, is that Frege allows for empty names at the cost of the Law of the Excluded Middle: a foundational principle of classical logical calculus, including Begriffschrift, in virtue of which the truth value of any well-defined formula is made determinate. Ergo, this is no good.

What Russell (1905) instead proposes is the new doctrine of logical descriptivism placing description at the core of reference as a phenomenon, even more so than the classification-oriented Begriffschrift. Descriptivism explicitly renders all object reference as description and I consider it more in chapter 3. What is important for present purposes is the motivation behind descriptivism vis-à-vis empty names, like the present King of France, against both Frege and competing analyses like Meinong (1905). The problem that Russell (1905) and Meinong (1905) – and other rough contemporaries, like Twardowski (1894) – are aiming at is what I will shortly be defining as the problem of *content* for object reference. From a broadly Fregean starting point they begin by accepting the idea that reference can be analysed through the mechanism of logical classification. In other words, that there is some information content aspect of reference (that Frege called the ‘sense’ or ‘thought’) complemented by a criterion of whether that information content is accurate (like the truth-value assigned to each thought, such that e.g. *a* is in

F or is not). They then ask what form that mechanism must take, in order to generalise more fully than Frege envisioned: in particular what form it must take to generalise over empty and non-empty names.

Otherwise put, the question was how to arrive at a theory of reference where every classification case for a label – including that there is no referent – can be given a general expression in the same logical language, consistent with the rules of the logical calculus that Frege had developed in the last century. I take this question to be the intuitive problem of (completeness of) information *content* for reference:

(C2) CONTENT

For some language $\mathcal{L}$ and set of labels L, I take the *content* of L to be the specification of all L in $\mathcal{L}$.

More than the previous definition, for coordination, the above leaves some room for unpacking but is ultimately quite simple a requirement. What it amounts to is a relationship between a set of labels that comprise the pool of entities whose capacity to refer we are interested in, and some arbitrary language (arbitrary in this definition) using which all those labels of interest may be specified without exception – so that for every label in L there is a mapping between that label and some appropriate formula in $\mathcal{L}$.

I then take the specification of each label in this manner to express its information content. That is, for my usage I will not delve deeper in what 'information content' ought to be, e.g. following information theory (Shannon, 1948; Mackay, 2003). My interest is rather in there being some general and uniform way to attach some kind of description *in one language* to all labels of interest in a theory of reference whereby the language will interpret or otherwise elaborate on these labels, for example by description. What is important to emphasise is that the sort of languages I will be discussing through the following are all interpreted logical languages per the template of a first-order predicate logic like Begriffschrift. And I do not intend to deviate from this idea of what specification languages are: were this the goal, a more general discussion of information content would likely be called for. Whereas what I am aiming to do with the above is not to define, or even prescribe features of the best sort of language for the job of specification. I am instead aiming to define what the job of specification *is* in a theory of reference: specifically, to attach some sort of appropriate information in a general way for every label of interest.

It should already be possible to see how Fregean semantics as I have described it is challenged by C2. Although it does account for the specification of each label, after which classes its referent belongs to, there is no general form of specification for there being no referent to classify. As a result, though it is ostensibly sufficient for explicit expressions like *Fa* presuming *a* is non-empty, the Fregean semantics I have described above fails to tackle cases where *a* is presumed non-empty *yet is empty* and so cannot be classified. What is useful to note is that the analysis explicitly given in Frege (1892a) is immunised against this concern, by limiting what labels are of interest. Effectively, Frege is saying that just labels known to refer, by some external means, and are thus validly *presupposed to refer* ought to populate L in the above definition. And this does solve the problem for Frege in the context of classifying reliable reports from a natural science already on its way to being perfected – to the extent of already knowing what referents there are to classify. But it does not solve the harder problem of providing a system for reference before the ontology of referents is even clear itself, and the labels may not refer to anything.

Another clarification is in order concerning labels themselves, which I have so far used in an intuitive rather than exact fashion. What I mean by 'label' in this context, and throughout the following, is that *token* to which e.g. information content might be attached, or a referent might be related, in actual use. That is: I do not mean some sort of abstraction, like the *type* 'Kripke' or 'cat' (which is also why I am avoiding 'name' for its similar connotations), ranging over all those signals that presumably pick Saul Kripke or e.g. all the members of the *felidae* family out from a bag. Rather, my intent is to analyse the *instances* of literal strings like ‘Kripke’ or ‘cat’ (or equivalent sounds or gestures), used in a particular context to ostensibly pick out a particular referent. The set of labels L is therefore populated by a vast number of duplicates, and a vast number of variations on the same name or description. Starting from those instances, I will (shortly) appeal to abstractions, but it is the instances that I analyse in this work.

It is lastly worth observing that I have been discussing specification *of labels* rather than referents, yet my previous definition e.g. of commensurable specification invoked specification *of referents*. That is because by the time a theory of coordination is desired, there is likely already a theory connecting any specification of labels to referents, by virtue of connecting the labels to referents. At any rate, as noted in the last chapter my interest here is in connecting these three aspects of reference (via the framework in chapter 5) under the assumption that they are all valuable – for the prima facie reasons I gave in my introduction and for methodological reasons I motivate in the next three chapters; so, the conflation is harmless so long as specifications of labels (i.e. content) have a reliable way to carry over to referents. Accordingly, my usage of ‘label’ to cover referring expressions of various forms – notably names and explicit descriptions – is intentionally ambiguous given the common treatment I will pursue for these.

These clarifications aside I now very briefly preview the Russellian solution to the problem of content before moving on to the final problem, of contact, common to both Fregean and Russellian semantics. The Russellian solution to the problem of content is – like the problem itself – inherited from Fregean semantics. Where Frege gave a system of reference just for non-empty names, Russell (1905) expands the same underlying general system of logical specification into the case empty names as per **figure 4**.

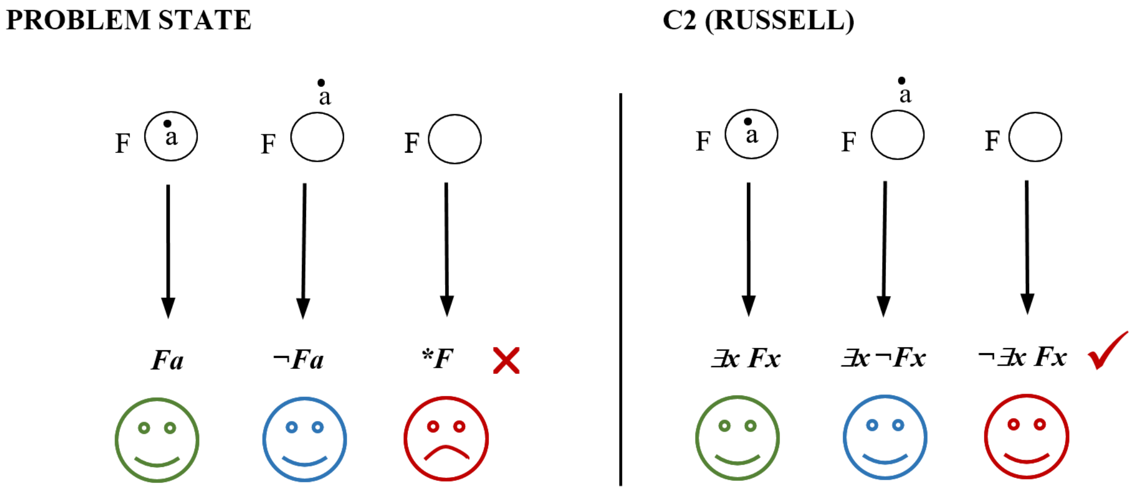


Figure 4. The empty name problem state (left) and the descriptivist solution.

Though I will not go into further detail about how the formalisations on the right-hand side of figure 4 work until the next chapter, the gist of what Russell (1905) gives here is a correct logical specification for every case, following which there is always a way to specify a label even when it refers to nothing. This is achieved by invoking the additional logical mechanism of quantification – developed by Frege in Begriffschrift but not previously used this way – to define three possible states against Frege’s two: *∃x Fx* (‘there is an *x* so that *Fx*’*)* classifying something as *F*, which we might then use *a* to symbolise; *∃x¬Fx* (‘there is an *x* so that ¬*Fx*’) classifying something outside *F*, which we can then again use *a* to symbolise; and also ¬*∃x Fx* (‘there is no x so that *Fx*’), indicating that *nothing at all is classified as F*.

In many ways this amounts to an upgrade of the Fregean system, making room for more types of label without significantly altering the premise that information content ought to be managed by a logic and the ‘purchase’ of this information content in the external, physical world ought to be managed by truth conditions for that logic. Using my taxonomy of empty names from §3iii all the cases of vacant names and mistaken names are accounted for; though with no way to distinguish between them except by the potential differences in truth conditions by context, which will amount to a modification of the system that I will discuss more in the next chapter. What this Russellian system, of expanding the logic while keeping much of the underlying framework intact, still does not account for is any inconsistent names.

iii. Sorting

To be clear, inconsistent names are not the only thing Russellian descriptivism struggles with. Yet it is sufficient to exemplify the sort of question motivating (what I take to be) the third part of the problem of reference. And that general sort of question is whether truth is sufficient to track what must happen when a label is connected to a referent, so that this connection is subsumed within a theory of content like the one in Begriffschrift, or Russell (1905) – or instead a separate analysis of contact is called for. That is: Fregean and Russellian semantics do include analyses of contact even though I did not cast it as such for Frege, and will only discuss it later for Russell. They account for when reference ought to connect (or not) to some appropriate structure via the logical mechanism of truth. Frege sets out what happens for cases where a known referent is added to a new class – such that *a* either is *F*, or is not *F*. And Russell aims to expand coverage to when the label being (in my terms) specified will not refer to anything as with ‘the King of France’, where there is no *a* to speak of in the logical level of discourse.

It is nonetheless important to have a well-specified goal for comparing Fregean and Russellian efforts against, even if (or where) these really are sufficient. That is what allows room for a separate question asked over contact. Not purely whether the first two approaches in the classical semantic literature are enough to address it but also what precisely they should address. Which also happens to be vital since neither Fregean nor Russellian semantics are sufficient to track relations between labels and referents; among other examples in the case of inconsistent names. I thus use inconsistent names to illustrate the broader issue I take Kripke (1980) to raise about reference. Namely that contact is a separate question.

The means by which inconsistent names pose a problem, for what was just overleaf a solution to what was just overleaf the problem (content) with Fregean semantics, and a means by which Kripke (1980) in particular addresses this new problem are both illustrated in the left and right-hand side of **figure 5**.

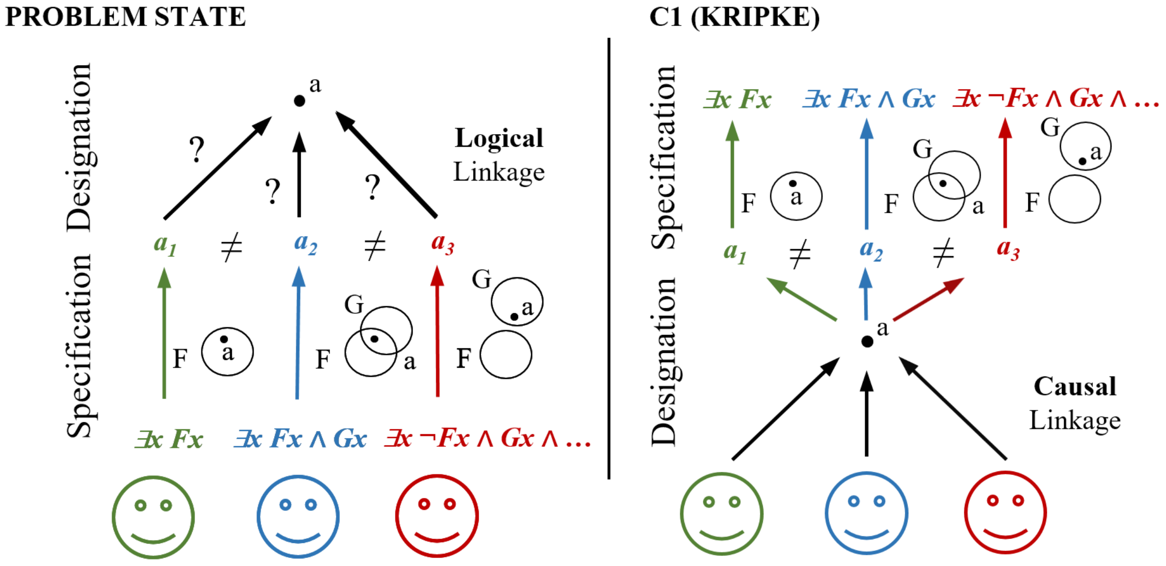


Figure 5. Logical (left) vs. causal linkage as the grounds for contact in the case of inconsistent names.

Once again the above is a prelude for a more extensive discussion in a later chapter and what I intend with it at present is to illustrate what the problem is and how two different solutions to it can function. The deeper nature, interactions and complications of these solutions are something I will discuss later.

What should once again be clear, though, is that the left-hand side is encountering a difficulty. In this case, how different logical statements, some of which (specifically relevant to the inconsistent names case) that cannot be true at the same time all nonetheless appear to connect to that same referent. The colour-coding for these different cases is to distinguish the cases rather than suggest they are different cognitive agents or 'private languages' (as with coordination) – for all intents and purposes these may be three different times, or other contexts, where just one cognitive agent refers to just one entity. And what is problematic is not synonymy, though that itself requires some elaboration for both Russell and Frege before him. What is problematic is that some of the information content attached to that referent is inconsistent with other *successful* instances of referring to a. (Note the non-italics, to mean the real-life entity a rather than its logical symbol *a*. This is the same a that is depicted as being in the set F in figure 1 above, where F is also not italicised for the same reason. I retain this convention throughout.)

Taking the inconsistent name example as the case in point, the cognitive agent in the example uses the label 'the man holding a martini' to successfully pick out the man in question, whom I will call Albert to make the example easier to follow (I will return to why that trick works so well in the next chapter). In that room on that day envisioned by Donnellan (1966) there is no man holding a martini, so what is attached to that label by its explicit language content is false. That is already a counterexample for the notion that truth is tantamount to establishing a connection to referents – since the information content of which truth or falsehood is logically asserted is false. To fully match the case in figure 5 it will help to then enhance the thought experiment, and imagine the cognitive agent looking for Albert is a friend of his with existing information content attached to his name, who knows he is looking for Albert, and so he is also using that information to find him. The information, among other things, includes the fact that Albert is teetotal and would therefore never be seen holding a martini or other alcoholic beverage.

The denouement of this somewhat convoluted drama is that if the cognitive agent uses the information they have on Albert to pick them out in the crowd, this will include that Albert is not holding a martini – and the cognitive agent can find Albert using information that this particular classification is part of. If the cognitive agent uses the label 'the man holding a martini' they will pick out Albert just the same even though that label explicitly attaches information content to its putative referent that is not true of Albert and (for the sake of the argument) no one else in the room as there are no martinis being served and even though the other way of picking out Albert comes with a logically inconsistent specification.

What Kripke (1980) concludes from this (as I will discuss in chapter 4) is that whatever it is that helps the cognitive agent pick out Albert both times seems unrelated to even explicit information content for the referring expressions (the labels) in question. Returning to that hard-and-fast distinction (from §3i) the label *works* in both cases even when the truth conditions imply that at least one of them should fail – and as a result there is a separate question to ask about when a label *works* distinct from information content. I encapsulate the call for a theory of 'getting there', or 'when it works', or *contact* as follows:

(C1) CONTACT

For some domain D and set of labels L, I take L to *contact* D, if and only if there is some criterion C based on which every member of L is assigned to a member of D or assigned to the empty set ∅.

That is to say: for that same set of labels L used in the previous definition (counting down to this one) I take a theory of contact to prescribe an arbitrary ‘criterion C’ based on which every label is assigned to an entity in a predetermined domain D (standing for e.g. the actual, external, physical world) or not. If the label is assigned to an entity in that domain then it has a referent and that assignment picks it out – and if the label is not assigned to any entity in that domain then the label has no referent. As for that information language in C2, the fact that the criterion C is arbitrary (in that it could be anything) does not mean it does its job arbitrarily. There are just a range of criteria that could fit this bill, and each of these would provide the basis for a theory of contact, as long as it sorts the labels into those two piles. For example: the Fregean and Russellian ‘Criterion C’ is *true description*. Where the truth value of a given description comprising the information content of a label is True, the label is sorted as having a referent by being specifically assigned to the entity in D matching that description. And Kripke (1980) can be taken to give a competing Criterion C based on different concerns whose assignment is a better fit for cases like inconsistent names and (as I will be discussing in the next chapter) insufficient labels.

4 Orientation

In all, through this chapter I have pursued a triple agenda. I established those questions that shape the specific sequence of philosophical ideas I take as my starting point: contact, content and coordination as separate yet interrelated concerns for the philosophical treatment of reference. I have also explored the Fregean project in depth as a whole, to understand and highlight what the philosophical treatment of reference should aspire to from his perspective: that puzzle-like analysis of all aspects of reference and their relations, including a possible reading of his analysis of thought leaning toward pragmatism, the idea that reference can capture an endpoint of the scientific process, and his generality most of all. All of these ideas recur in the following chapters even besides coordination, often in surprising places.

Lastly, through this tour of mostly problems and questions I have attempted to lay out the majority of tools and theoretical ‘dramatis personae’ from the philosophical perspective. Labels, designation and specification, my taxonomy of empty names, and several other minor elements and clarifications will all contribute for making the philosophical work involved in the next three chapters easier to develop.

I revisit Frege for one more detail to do with intention and usage at the outset of chapter 5. I return to Russellian descriptivism overleaf for the next chapter. And I return to Kripke across both of the next two chapters first for his deconstruction(s) of descriptivism, and then for his causal theory of contact.

Chapter 3: Content

1 Meaning in the Head

> “Cut the pie any way you like, 'meanings' just ain't in the head!” (Putnam, 1975, p. 227.)

In the previous chapter, I sought to establish the basic questions asked during a defining debate for the modern philosophical study of reference, and relate them to the notion of a triple problem of reference encompassing the discrete (but interrelated) elements of contact, content and coordination. During this overview I considered what each of these parts provides for a theory of reference, in the context of the problems it can clarify by its separate treatment: the sort of problems that really only concern a theory of contact, or of content, or of coordination largely discrete from the other two. In the present chapter, based on this analytical separation as I have advocated it so far, I move to consider one of the parts in isolation: the problem of content, its most influential solution and the ways this solution is challenged.

The problem of content as I have cast it so far concerns the intuitive aspect of reference wherein some sort of ‘meaning in the head’ is part of the process. Some internal state, or concept, underwriting what the label being used really has to say about a referent; in my terms, the information content it attaches. Such a notion of a ‘meaning in the head’ itself is, per the famous quote the shorthand comes from, not an obviously accepted one in the philosophy of reference. In particular, semantic externalism, which I will discuss more across the next two chapters and work like Putnam (1975) is a notable proponent of, denies the need for content (as I have understood it here) to participate in the analysis of reference. At the very least, considerable doubt is cast on the suitability of logical description for playing a lead role – if it plays any role at all – in the analysis of reference by several influential counterexamples, one of which I have already presented in the previous chapter under the category of ‘inconsistent names’ and the rest of which (at least the rest I mean to consider) come from one notorious source, Kripke (1980).

My work in this chapter will thus be threefold: firstly, I present Russellian descriptivism as a theory of content, and specifically the theory of content I will by and large accept and defend through this work. Secondly, through the prism of the famous counterexamples, I consider why the theory was rejected – and eventually ejected – from the analysis of reference. These two goals occupy the first main section below and effectively complete the philosophical presentation I briefly started in the previous chapter.

My third and more ambitious goal is to then examine how one influential way (empirical) psychology has come to understand ‘meaning in the head’ may complement and in many ways safeguard the view of content inherited from Russell, at least as I will have laid it out. In this latter main section, I take in selected results and ways of thinking about conceptual cognition, and their relation to the information content of reference. In particular, I focus on the understanding of concepts as hypotheses of what the external world contains rather than representations of sense data. At the close of the chapter, I last but not least consider *labels* as objects of analysis for cognitive psychology and their exact link to content.

2 Descriptivism

i. Form

In this first section on the use of logical descriptions as a theory of content Bertrand Russell inherited, expanded, and (subtly) realigned from the Fregean system of meaning, I return to the illustration from the previous chapter of the fundamental problem this novel analysis was designed to solve. At the end of the section my aim is to have made clear what exactly has been done to the Fregean analysis on the left-hand side of **figure 1** to transform it into the Russellian analysis on the right-hand side – and how the Russellian analysis works in its formal capacity. I then consider its broader philosophical purpose over the following section, before moving on to the most famous foibles of its generalised application.

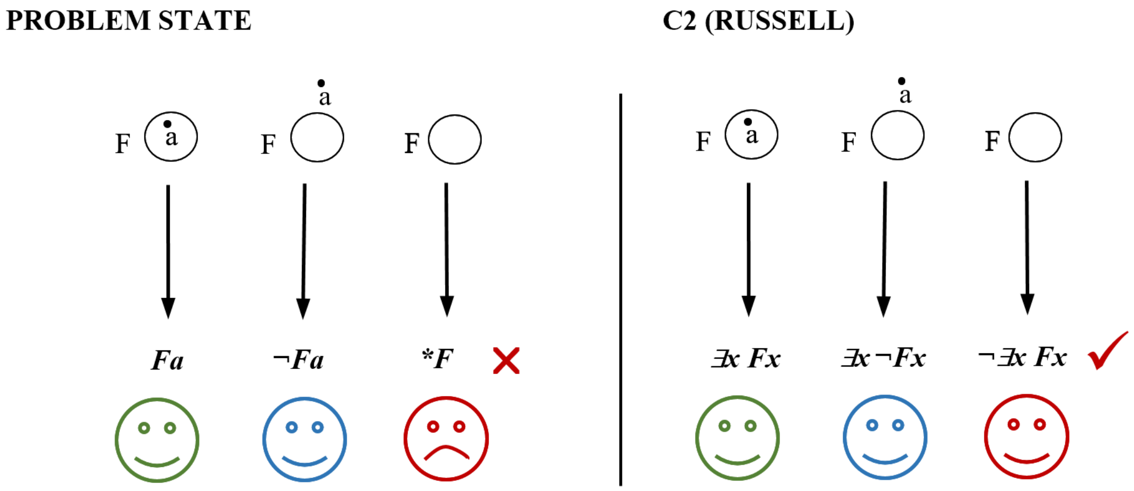


Figure 1. The empty name problem state (left) and the descriptivist solution.

The more basic Fregean form of information content serving as the baseline for this transformation is the saturated function-argument pair I have discussed at some length in the previous chapter. In every explicit conjecture expressed by a Fregean thought there is some sort of object that is being classified. The object being classified is represented by an object in the logic of thought, such as *a*, and the class relative to which the object is being considered is represented by a predicate in that same logic, like *F*. As a result, the conjectures relative to *a* and *F* can take only two forms, that *a* is in *F*, and that *a* is not in *F*. These are represented by *Fa* and ¬*Fa* respectively and form the bedrock of both Fregean content (again per my sense of content; the information attached to some referential expression), and the logic itself that expresses it. The pair of cases exemplifies an axiom called the Law of the Excluded Middle. And for Frege (1892a) these logical expressions, by formalising the information content attached to an otherwise 'noisy' entity formulated in natural language or even mathematics, make meaning amenable to truth-functional analysis. In other words, they put the information content of referential expressions in such a form as to allow truth to then support a theory of designation, or contact: where the formulae behind labels (as I have called any sort of referring expressions involving objects) are true, these offer a window into the nature, or at least the properties of *a*. And in the completed logical language, where all conjectures have been tested there is a unique sequence of classification giving every property of *a*. Finally, in one reading of the Aristotle footnote this also gives the information content of *a* simpliciter – in other words the information content of just encountering the label *a*, with no conjectures attached.

Yet as I have discussed in the previous chapter and will only briefly retread here this analysis does not suffice even for the idealised case of the completed logical language where the complete classification of *a* is available from just its label, and thereby perfectly captures its properties. The reason is that this will not include the class of labels with no referent whose properties to capture. And though this too is to have been eliminated in the completed logical language (cf. Frege, 1892a), unlike putative analyses for the information content of (just) *a* where it does designate a referent but just *some* of its properties are known, which Frege supports and endorses with both the Aristotle footnote and his later work (cf. Frege, 1918), there is no putative form of there being no referent. There is an analysis of omitting the requirement of designation altogether, for fiction, which I once again will not delve into here, but the option of dropping the referent as an outcome of classification is not logically 'put on the table' at all.

In response to this problem, and its consequent paradoxes for e.g. the Law of the Excluded Middle by trying to assign all non-designating labels (i.e. empty names) falsehood via the empty set as a stopgap (albeit a stopgap some authors appreciate; e.g. Beaney, 1997), Russell (1905) suggests a modification: where Frege started with primitive statements like *Fa* linking a functional expression *F* to a constant *a* – so that there has to be some sort of *a*, for the expression to work – Russell replaces the constant with a variable *x* such that the primitive expression of interest for the analysis of reference is instead $\exists x\, Fx$. In plain English the basic classification statement allowed by Russell (1905) is that there is something which we classify under *F*. And when for example a is in F (as above), this is true: there *is* something classified as *F* though specifying it to *a* requires logical clarification which is not important at present.

The real value of using a variable for this primitive form, as opposed to a constant, is in distinguishing two further cases (vs. just one) for the classification of *something* relative to *F*. First, that familiar case from above where a is not in F – so that *x* could be replaced by something explicitly not in F available for classification elsewhere: an object e.g. called *a*, whose classification relative to *F* is that it is not in it. And second, the previously problematic case. In this final option – leftmost in figure 1 – *there is no object in F* the variable *x* could be replaced by, so there is no object in F at all and *F* is an empty class. The upshot of this new option is resolving ambiguous classifications, like 'the King of France is bald', where the fact of the matter is not that the King of France (at the moment) is *not* bald but that, instead, there is no King of France to speak of. The object which must be King of France as well as bald is not bald, because there is no object to classify by meeting the initial criterion. Nothing is substituted for x.

As a result of this upgraded form of logical description Russell (1905) explicitly advocates what was only vaguely gestured at by Frege (1892a): that every possible referring expression, whether general or specific, empty or non-empty, could have its underlying logical content expressed via description. For general terms like 'cat' the description is something like the above, $\exists x\, Fx$, or 'something is a cat'. For specific terms like 'the King of France' the description is qualified, e.g. as $\exists x\, Fx \wedge (Fy \rightarrow y = x)$, or, 'something is the King of France, and every logical symbol for that entity marks the same object'. And empty vs. non-empty names are distinguished, not in advance like Frege (1892a) had suggested, but as part of the same process of conjecture whereby *Fa* is tested and found true or false in Fregean reference. Whereas Frege only tested membership Russell (1905) also uses truth to test for *existence*.

And since Russell (1905) is advocating that every possible label may be analysed by description this way, he also requires this of proper names like 'Napoleon'. Some sort of description should exist, he reasons, so that 'Napoleon' reduces to that description like the more explicit labels ('King of France') and as a result the descriptivist project he spearheaded could achieve complete generality of coverage. Taking this one step further, he prescribes that "denoting phrases [i.e. labels] never have any meaning in themselves, [instead only] every proposition in whose verbal expression they occur has a meaning" (Russell, 1905, 480). In other words, labels should be replaced by an appropriate logical description – and that appropriate logical description should be the only target of analysis for a theory of reference. In the following section I will now aim to unpack what led Russell (1905) to impose this requirement.

ii. Function

So far I have discussed the updated form of logical description that Russell (1905) advocated to push the Fregean presentation of information content into a higher tier of generality. And I have also noted the proposal accompanying his update, that every instance of reference is a description and as a result the analysis of reference is 'just' the analysis of logical description. I now explore what this buys him: what kind of picture this reductive approach to reference motivates, set apart from Fregean semantics.

To start with, this idea that the analysis of reference is the analysis of description is not a far cry from the Fregean project that preceded it. Frege (1892a) would be at home with exchanging 'thought' with 'description' for the previous sentence. And despite his pains to distance himself from that 'sense and reference' analysis he takes his system to be replacing, there is an argument to be made that Russell is only exchanging the trappings of 'sense and reference' that amount to whether a logical description is true, with the more deflationary trappings of 'denotation' that amount to whether a logical description refers by virtue of its truth value – although I will not pursue this specific exegetical argument further. Rather than its trappings, what I will focus my discussion on is the use case for descriptivist reference not just as a more inclusive model for the form of the information content of referring expressions, i.e. labels, but as a different model for what that information content actually represents, relative to Frege.

Specifically, and much like Frege before him, Russell (1905) did not intend his theory of reference as an intellectual exercise. Just like Frege there was a particular sort of activity he intended his theory to help set the right goals for, and just like Frege that activity was natural science: the discovery of what objects belong to what categories and (in the descriptivist case) the indexing of what objects there are. As made very clear in his subsequent contributions to a joint form of logic-*cum*-epistemology Russell (1911, 1917) intends his descriptivism to index the alignment (or not) of human sense data to physics. It is this philosophical 'master plan' just as much as the form in which it is delivered that makes what Russell had to say about reference overall remain vital to a notion of content as 'meaning in the head' – and by extension to my present analysis, both of content, and of how content and contact are linked. To better understand the viewpoint informing his master plan I will now therefore offer a tour similar to the one for Frege and the thought, in the previous chapter: a bird's-eye-view of what was intended, even if it was not entirely plausible or wholly realised, to be the 'completed' Russellian descriptivism.

At the starting point for this tour there are two discrete sets of entities, the *physical objects* populating the world and the *sense data* available to a putative human observer – the kind of embodied cognitive agent I have been assuming as the end-user for the mechanism of reference. Unlike Fregean 'thought' these sense data are not posited at any sort of inherently logical or transcendental level of description. They are what we collect when we see, smell, hear, etc. so the 'mission' is not to discover them in the wild as we would the (correct) Fregean objects and classes (cf. Frege, 1918). This mission is different:

> "Physics is said to be an empirical science, based upon observation and experiment. It is supposed to be verifiable, i.e. capable of calculating beforehand results subsequently confirmed by observation and experiment. What can we learn by observation and experiment? Nothing, so far as physics is concerned, except immediate data of sense: certain patches of colour, sounds, tastes, smells, etc., with certain spatio-temporal relations. The supposed contents of the physical world are prima facie very different from these: molecules have no colour, atoms make no noise, electrons have no taste, corpuscles [sic] do not even smell. If such objects are to be verified, it must be solely through *their relation* with sense-data: they must have some kind of correlation with sense-data and must be verifiable through their correlation alone." (Russell, 1917, p. 113, my emphasis.)

Which is to say: we start with one set of entities, sense-data, and the end goal is another set of entities, the objects of physics. And the problem we are facing for Russell (1917) is not *locating* the sense-data in the external world, as was the case for the objects of thought and the physical objects they stood for in the Fregean project. The problem is instead *relating* the sense-data successfully to the second set of entities, the contents of the physical world, as the objective for a successful natural science. Explicitly, Russell (1917) sets the agenda for his descriptivism as the connection of these two sides, the proximal and the distal (more on this particular terminology in the next chapter), by way of logical descriptions.

It is important to make clear at this point that the descriptions are not of the 'underlying entities' at all. As far as Russell is concerned, these descriptions are of the pattern of sense-data associated with what physical entities whose properties we hope to 'verify' rather than that pattern of properties themselves – such that the slogan behind his project is 'the relation of *sense-data* to physics' (Russell, 1917). And though much of his analysis is unnecessary to go through, in light of what I discuss in the next section and overall, it is critical not to lose track of this distinction. Russell advocates the relation of one set to another entirely, through the medium of logical description, where Frege advocated matching contents of a single set instantiated 'in the head' to their counterpart instances instantiated 'in the world'. There is no ambition on behalf of the Russellian descriptivist to capture the external world directly. The only ambition is to capture the patterns of sense-data the external world causes and thereby infer its entities in the form of the descriptions (or rather the true subset of these descriptions) that sketch their content. Put differently, Russell never intends to substitute some inferred entity *a* for *x*: his perfected science is one of true descriptions where an entity *could* be substituted in principle but he does not consider such a substitution epistemologically licensed, because we have never seen nor could ever possibly know a. (Note again my non-italicised usage. What I mean is Russell does not think the real-world entity a can be reached except by its 'residue' on our senses, whose patterns we collect, so a description is as deep and as informative as we can go. I set aside his comments on the sense-data of others for this context.)

What these descriptions *do* is therefore collate sense-data, and just like Frege they are implicitly taken to be the content of thought – or at least perceptual cognition – just as much as they are the content of linguistic reference. As Russell (1911) clarifies in a different essay linked to the one where he sets out the descriptivist theory, the (in my terms) information content of both perception of and reference to a putative external entity is a case of 'knowledge by description' in lieu of 'knowledge by acquaintance' where the latter is impossible to acquire (past exotic circumstances not related to my present analysis). All we have in both cases are indirect evidence for entities, which Russell (1911) takes perception and reference to collate in the specific form of these descriptions I have considered in the previous section.

For example, 'Walter Scott' reduces to the information available about Walter Scott, in which Russell (1905) does not include his name to avoid circularity (i.e. Walter Scott refers to the man called Walter Scott appears circular to Russell, as indeed to most philosophers, though I will argue against it below). His completed description is some sort of (long) list of everything Walter Scott can be known to have done, been and looked like, 'grounding out' at the level of whatever sense-data are observable around Walter Scott.[13] This would include information over what he looked like and, as Russell (1911) allows and specifically anticipates, information of what sort of testimony we have recorded people as saying; so one part of Walter Scott's description could be that he was *reported* to have written a famous book – literally that someone said this of him rather than whether or not he did it, in this particular instance. Likewise, the breakfast item known as a 'bagel' is a collection of sense-data available around a bagel, so far as the human capacity to perceive it is concerned; and using the term 'bagel' as I have here will contain that very same information (in a perfected science) as is involved when the bagel is perceived. Clearly the capacity of the labels 'bagel' or 'Walter Scott' or of more explicit descriptions that are still

[13] NB 'observable' here. Per Russell (1911) the components involved in these descriptions need not be observed as long as they are observable: i.e. as long as they are potential sense-data, and therefore the right type of 'stuff'.

not couched in terms of low-level sense-data, e.g. ‘the King of France’, to encode information like the above is subject to the knowledge of the user. Which is why Russell (1911) in particular is taking this analysis to be as much epistemology, the theory of what is known, as philosophy of language or logic.

As a result, Russellian descriptivism aims to offer a general, regular form for all labels, in the form of the underlying descriptions (often partial) couched in terms of observable rather than inferred entities, not just to make content clear but to clarify the way perception and reference alike give us knowledge. This particular part of the overall endeavour is easy to overlook given some of its formal deficiencies; and the fine detail of how exactly these descriptions can be tested against future observation is largely left for the reader to imagine in the sources I have discussed above. Nonetheless, there is some sort of picture for why these descriptions are inherently important emerging from the above considerations – not just for the study of reference as a linguistic phenomenon but for its value in acquiring knowledge. When it is (somehow) tested that a ‘bagel’ is a certain sort of thing, e.g. $Fx \land Gx$ (I omit the existential clauses) rather than some other sort of thing, e.g. $Fx \land \neg Gx$, then some sort of understanding of what a ‘bagel’ *might be* is reached, where descriptions encode the development of this indirect understanding.

Even more specifically, what Russell (1917) takes to be the basis of that understanding is exemplified by our (in his terms) knowledge of the real shape of an object by combining its perceived perspectives – the actual and only data connected to that object that we have access to. In compiling our sense-data overall, Russell argues, we may progressively identify a regular pattern of sense-data being connected to a particular (putative) point in (putative) space. From this, we build *a descriptive theory* (my terms) of what occupies that space, and eventually infer the *cause* of that sense data we have been collecting:

> “In physics as commonly set forth, sense-data appear *as functions of physical objects*: when such-and-such waves impinge upon the eye, we see such-and-such waves [and colours] impinge upon the eye, and so on. But the waves are in fact inferred from the colours, not vice-versa. Physics cannot be regarded as validly based upon empirical data until the waves have been *expressed as functions* of the colours and other sense-data. […] We have therefore to solve the equations giving sense-data in terms of physical objects, so as to make them instead give physical objects in terms of sense-data.” (Russell, 1917, p. 114, my emphasis.)

Russellian descriptivism is thereby as much a theory of inference as it is a theory of reference; and the device of a descriptive theory used to infer the causes of proximally-available data will return below – though the content for that descriptive theory will admit an alternative option Russell did not envision.

I have not discussed the role of truth in the above except to suggest true descriptions imply knowledge for two reasons: firstly because nothing has changed in this regard from the Fregean emphasis on truth as the decider of classification, though Russell does have an elaborated theory of truth (Russell, 1906) and secondly because, as I will immediately move on to discuss, the notion that true descriptions track referential success (i.e. contact; i.e. designation) stands on some rather precarious ground. So I present the above as a clarification of the role of information content, and the insistence on substitution of that information content for whatever label is provided as a result of the former doing all the work, in what Russell (1905, 1911, 1917) has endeavoured to build overall. I do not intend to justify the full package of views that accompany his analysis of content, except where these relate to it as a descriptive theory.

iii. Foibles

Having spent some time on the rationale behind the descriptivist reference Russell (1905) advocated – to supersede Fregean semantics, and to elaborate on the connections between information content and knowledge – my focus now shifts to the most prominent reasons his attempt did not (overall) succeed. My exact interest here will not be to contest the overall conclusion by Kripke (1980) that descriptions are not sufficient to establish contact between a label and its referent. Rather, I will aim to catalogue a core set of foibles for descriptivism that Kripke (1980) and Donnellan (1966), and further members of what was at the time the 'descriptivist programme' (e.g. Searle, 1958; Strawson, 1961) have identified that concern the feasibility of descriptivism even in its central role of cataloguing information content. My reason for doing so will be to then contrast these complaints with the 'empirical descriptivism' of psychological science, to explore the extent to which it addresses them, if taken as a theory of content. The agenda is therefore once again, as with the last chapter, to collect problems in search of solutions.

The challenges facing a descriptivist theory of content (and by extension a pure descriptivist theory of reference) that I will consider here are roughly divided into three categories: cases where a description is either *implicit*, or *explicit but insufficient* to account for the information content apparently encoded, cases where an explicit description is *inconsistent* with properties of its referent but nonetheless works the same as a consistent equivalent, and finally cases where *context of use* alters the referent where the description provided otherwise remains constant. This last case can itself be divided into weaker types of context effects, like the ones from Donnellan (1966) and the 'vacant name' case I considered in the last chapter, and the stronger, modal context effects whose impact principally concerns Kripke (1980). I will argue here that an 'empirical descriptivism' can account for all but those stronger modal context effects, following a more empirical approach to information content over the remainder of the chapter.

The first item on this list, cases where descriptions seem insufficient or indeterminate, is arguably the most intuitive. Descriptivism as a whole is elaborated by Kripke (1980) as a link between belief in an arbitrarily large set of properties in the form I considered in §2i, and a putative entity which has these properties. Yet the question of how large that set must be, or what proportion of the true properties of an entity that set must capture, is left uncertain by the (endpoint-oriented) definitions of descriptivism. Russell (1905) makes it clear enough what a logically adequate description ought to cover. Taking the most specific use case of this system, descriptions picking out exactly one thing, it will be a collection of those 'identifier properties' that one entity and only that entity satisfies. In other words, a bundle of properties exactly one entity has in common. If there is one King of France then the description $\exists x\ Fx$ is sufficient for the label 'the King of France' (or his name, that Russell takes the minimal description to stand for), where 'F' stands for being King of France. The full span of information content that *can* be attached to the label is greater, and includes all the properties the putative entity may be associated with through observation. But the set of properties necessary for the description to be a description of a unique entity is just the minimal set by which a unique entity will be selected. By contrast, the label e.g. 'the scientist in a wheelchair' when considered to be an explicit description is not such a minimal set. There are doubtless numerous scientists in wheelchairs, and yet the label seems sufficient to track precisely one scientist in a wheelchair (until his recent death), namely the physicist Stephen Hawking.

Along a similar vein, knowing just the name of an entity is often enough to select exactly one entity in the world by that name even though there are plenty of other (known) candidates: in practice, 'Trump' selects exactly one real-estate mogul and US President, in spite of several other people by that name – including his son – that someone may be familiar with. 'Trump' suffices to select one man where this should not be possible purely based on the explicit information it encodes. And the means the implicit information associated with 'Trump' appears to involve being the father rather than brother of Ivanka Trump – in other words the *implicit* information content attached to the label 'Trump' – is mysterious. These cases suggest that set-based identification is an ill-fitting record of what information content is actually attached to a label, even setting contact aside, and that links between labels and their implicit descriptions (as in 'father of Ivanka Trump', vs. 'brother of Ivanka Trump' or indeed 'Ivanka Trump' herself, for 'Trump') are, again, insufficiently given by the property-set-based criteria Russell (1905) exclusively considers. The set of properties needed to identify an entity appears 'beaten to the punch' by much weaker sets – both for determining the referent and for determining the information content attached. Whoever 'Trump' *really is* (contact) their implicit description includes 'US President', and not 'brother of Ivanka Trump'. And the exact span of that implicit description is also uncertain: what exactly is and is not contained in the set of properties attributed to *some* entity via the label 'Trump'?

The opposite problem is inconsistency, between the logical description attached to a label and the real description of its referent based on the Russellian system. That is: assuming some description $Fx \wedge Gx$ is attached to a label, this problem arises when the full (scientific endpoint) description identifying the intended referent includes (e.g.) the fragment $Fx \wedge \neg Gx$ that is logically inconsistent with the previous formula. The putative description says some entity is in G. The genuine description says it is not in G. Russell (1905) would have a very clear response for this case. The description used is false and hence the putative entity does not exist because the one that does is not in G. As a result the former is empty. In practice, however, the violation of one clause does not appear to cancel the attempt or the outcome: explicitly describing someone *very much like* Aristotle (to use the standard example) can often enough allow someone to understand the description as attaching information content to Aristotle even if parts of that information are incorrect; e.g. the colour of his sandals. As Searle (1958), and Strawson (1959) hot on his heels, consider from within the descriptivist programme, there would appear to be elements of the logical descriptions attached to labels whose importance is not as critical as others. There would e.g. be no problem with identifying the ancient philosopher born in Stagira who trained with Plato and taught Alexander the Great *and hated bees,* even if Aristotle loved bees. Furthermore, this tolerance is extended to attaching novel information content to what is otherwise a false description: in saying 'the ancient philosopher born in Stagira, who trained with Plato and taught Alexander the Great, and hated bees wrote the Prior Analytics' the information content associated with Aristotle is suitably enhanced.

As one can get Aristotle partly right with a description, and that description would still 'survive' as an attempt to identify the referent, or to add additional information content to what is already available, it is difficult to see how a simple 'yes or no' criterion like the one in Russell (1905) is the right approach – again not just for the link between description and contact, but also for the link between descriptions and other descriptions, where the latter are intended to elaborate on the same referent whether or not it was determined by description in the first place. From a logical standpoint, a wrong description might function like *an equivalent that was not wrong* with respect to its false classification choices, so that it still offers information on Aristotle. And so at minimum the suggestion is made, by a range of authors considering this problem from within, to introduce a statistical element of e.g. taking an average of the right vs. wrong classification decisions, or weighting them as appropriate (Strawson, 1959) – or at any rate that "the logical nature of the connection of such characteristics with the man's identity may again be loose and undecided in advance of dispute" (Searle, 1958, p. 172). Ergo to loosen up descriptivism.

The last wave of criticism corresponds to the context-related concerns, some of which I have raised in the previous chapter. Firstly, the case where descriptions are made true or false by the passage of time (or alternative 'natural contexts' as I briefly called them) which is not very difficult to address even in the Russellian framework – although it does require explicit use of time-, space- and (overall) context-capturing predicates to render correctly. I will not dwell on these issues for now past my definition of the class of 'vacant names' in the previous chapter: the labels whose referent is unavailable but could be 'found' somewhere, as opposed to 'mistaken names' whose referent was never available anywhere. I return to this element in chapter 5. Secondly, the more problematic cases raised by Donnellan (1966) but also Strawson (1961) and Evans (1982) and others, where context of use seems to alter the ruleset for assessing descriptions. Once again this happens by ignoring false classification, albeit now by context rather than convention, as I have already considered in the previous chapter for 'the man holding a martini'. Here the problem seems to be not that the description is understood to be false and ignored

– setting aside the version in the last chapter where the man was known to be teetotal – but that a false description works in context, when what is otherwise false about that description (holding *a martini*) does not matter to that context. This suggests context could render the truth of descriptions irrelevant.

The typical responses to these problems from the descriptivist camp are to divide descriptions into the sort of description intended to define, add to, or otherwise modify any information content attached to a referent – what Donnellan (1966) calls the 'attributive' use – against the sort of description intended to select some suitable entity by approximately capturing its qualities (much like Aristotle above, too) – what Donnellan (1966) calls 'referential' use. For my purposes, such distinction will not be needed: what is important to take home from these cases is just that contexts do appear to modulate reference. As I will argue below, there is no need to split descriptions by purpose, where the context can instead be directly involved in the specification of information content – so 'martini' is false but effective not because it is a type X rather than type Y use of descriptions but because 'water' and 'martini' literally mean the same thing here. To be clear: such a solution would also constitute a departure from Russell (1905) where there is no context factor included in descriptions. This is a departure packaged into the way that I will suggest empirical psychology can effectively complement the descriptivist programme.

There is nonetheless one last sort of context effect that cannot be ignored, or analysed away, as easily. The context in question is modal context: those 'what if' cases where reference seems to be preserved against all predictions or prescriptions from descriptivism, laid out by Kripke (1980) and signalling an end to the reign of descriptivism as the default analytical paradigm within the philosophy of language. These convincing counterexamples challenged (among other things) whether *any* part of a description at all can be privileged over others, and whether instead *no* part of a description is essential to it at all: what if Aristotle was a woman? What if Aristotle was black? What if Aristotle was not born in Stagira and did not learn from Plato and did not teach Alexander the Great? Kripke argues these questions are never such that by asking them, one is asking some equivalent of 'what if Aristotle was not Aristotle?' Instead of modifying his definition, they all modify the (in my terms) information content of whatever hypothetical character is being considered in contrast to 'our' Aristotle, who always remains the same. That is to say: in considering variations of entities across modal contexts, changing descriptions fail to affect our original entity (whose variations these are) entirely because any property could be denied as part of the 'what if' story, without harm to the version of Aristotle we start with. More than proposing descriptions per se do not fix referents, Kripke (1980) argues reference does not depend on them or on any notion of information content at all for its primary function I have termed contact. Something else fixes the referent of a label to which descriptions are at best secondary as I discuss in the next chapter.

Kripke (1980) thus contends there is no logical property, attributing sense-data or observed properties to an entity like Russell (1905) had in mind, that we cannot deny. And so a description cannot be how we are keeping track of what 'Aristotle' refers to, while in the what-if process of denying its elements; or when a description is completely inaccurate that no part of it is right, but the inaccurate or impostor entity so described remains the apparent target of a label, as I will be considering in the next chapter. I will sustain the objection that description is insufficient for contact without much debate. Yet it is also important to stress that Kripke does not deny the role of descriptions in providing information content for labels. On the contrary, once the element which does set the referent has done the job, information content attached to that referent may be expressed through description. Kripke only reverses the order of effect such that contact is resolved first and description second. He does not eliminate descriptions. Both because descriptions are thus still allowed even in light of this much more serious objection and also because I intend to argue against this strict ordering, it therefore remains important for me to call attention to one place where (a type of) descriptivism is alive and well and able to deal with all of the above challenges except the modal one. And this descriptivism can still pull a lot of weight as a result.

3 A Conceptual Solution

i. Descriptions

So far in this chapter I have considered the descriptivist system of reference: a philosophical attempt firstly to regiment and formalise any information content involved in reference, secondly to establish the ‘description’ of referents as the sole mission of a theory of reference in general. As I do generally agree with Kripke (1980) about the most serious limitations of descriptivism as a theory of contact, it has been my focus to explore how descriptivism can be challenged even as a theory of content per se. Focusing on just the core of this literature I have tried to raise a list of aspects in need of clarification: cases where descriptions seem mysteriously implicit, or explicit yet insufficient, cases where they are inconsistent with known properties of what they describe yet nonetheless freely replace the consistent version, and cases where context (all other things being equal) appears to invalidate their information.

Setting my sights on those points of weakness for that purely philosophical approach to descriptivism, my next move is to now consider a psychological approach, which I will argue matches descriptivism in form and function as (at least) a theory of information content for reference while avoiding its most obvious foibles. The psychological approach to ‘meaning’ as a whole is vast even for the minimal sort of application to (just) labels specifying objects I consider here; and my intent is not to review or even consider the full span of psychological theories that can pass muster as theories of information content for reference. Instead, I will first consider basic ways that ‘concepts’ studied by psychological science carry and organise descriptive information and then consider *one* theory of information content ‘in the head’ from the recent literature. This theory casts concepts as (descriptive) reconstructions of inferred entity classes, rather than descriptions of the evidence for that inference (like sense-data), and thereby satisfies the form and function of descriptivism in a different way than Russell (1905) had envisioned. The particular premise, that a descriptivist theory of content can be based on what is ‘out there’ rather than what is ‘in the head’ or ‘given by the senses’ will then form one part of my suggested framework promising a more complete solution to the problem of contact, content and coordination for reference.

For the moment, I begin with the most basic question asked by psychology over ‘meaning in the head’ that is also of immediate relevance to presenting a conceptual theory under the guise of descriptivism. That question is whether there is such a thing at all as description based on properties, implemented in humans. I am importantly *not* asking or considering at all what the language of that implementation is: whether a symbolic ‘language of thought’ (Fodor, 1975; 2010), or a distributed connectionist network (Rumelhart & McClelland, 1986; Rogers & McClelland, 2004) or a dynamic system (Thelen & Smith, 1994) and/or affordance-based radically embodied variant thereof (e.g. Chemero, 2009) is responsible for *representing* the sort of descriptions and the sort of categorisation and inference mechanisms from which I draw any of my conclusions here *in the brain* is not something I will be considering further. If there is one representation-related assumption I am making it would amount to requiring some type of structure that can be functionally interpreted or explained by a higher-level theory on the organisation of and inference from object-related information, using postulated theoretical entities I call ‘concepts’. That is the extent of my representational commitment: X, where X may implement the desired theory. Danks (2014) calls a version of this (non)commitment ‘representational realism’, where the only sort of thing assumed is that representations invoked by a theory find consistent purchase at other parts of the same level of description; in the sense of Marr (1982) allowing for a phenomenon to be described at a sequence of levels increasingly sensitive to physical (vs. functional, algorithmic) implementation.

Having made my representational non-commitment as precise as I can, it is now less egregious to say (bearing in mind the previous) that an important part of meaning in the head does involve description.

Perhaps the best and most theoretically non-committal illustration of this phenomenon is by examples from clinical neuropsychology where the standard operation of the brain (and therefore cognition) is disrupted by physical trauma. The approach to scientific psychological explanation underwriting these examples is *dissociation* (Shallice, 1988): selective disruption of some behaviour, against preservation of other behaviour. The contrast of some mechanism that no longer works, against something that still functions (broadly; cf. McIntosh, in press) as intended, licenses inferences structurally dissociating the two; with various gradients whose strongest version is the so-called double dissociation where each of these mechanisms is found to function in the absence of the other (though e.g. Van Orden, Pennington and Stone, 2001 represent the more sceptical wing of psychological science concerning dissociations).

For example, a double dissociation for a patient unable to speak (Broca, 1865) against another patient unable to understand language (Wernicke, 1874) was influential in suggesting that the production and comprehension of natural language have separable mechanisms (Lichtheim, 1884). Though even that most famous of monolithic neuropsychological findings is under doubt, e.g. by Fridriksson, Fillmore, Guo and Rorden (2015), over the course of time since the 19th century developments and observations in neuropsychology have provided robust enough behavioural evidence of the underlying mechanisms behind several cognitive features, most notably perception, memory and language (see e.g. Andrewes, 2015). More relevant to present purposes are the several ‘blockbuster’ case studies of patients who are challenged by attaching any sort of information content (per my terms) to entities in the world around them, and/or to their verbal labels (Warrington, 1975; Hodges, Patterson, Oxbury and Funnell, 1992): “Have you ever been to America?” “What’s America?” “What's your favourite food?” “Food, I wish I knew what that was” are typical exchanges with the latter sort of patient (Hodges et al, 1992, p. 1786). Cases like these have motivated the notion of a ‘semantic memory’, to mean a memory for conceptual information as opposed to events – a memory for “words and other verbal symbols, their meaning and referents, about relations among them, and about rules […] for manipulating them” (Tulving, 1972, p. 386). In other words, a type of memory storing precisely that sort of thing that I have called ‘content’.

Work on semantic memory mainly stemming from neuropsychological evidence like the previous has ranged far and wide in the years since its establishment as a construct long after the time of Frege and Russell had passed. Out of all this work, recently summarised e.g. in Yee, Jones and McRae (2017) or with a little more detail in Yee, Chrysikou and Thompson-Schill (2014), what I select here is what has most significance for establishing whether, and to what extent, semantic memory involves description. Which is to say: does the logical structure $Fx \wedge Gx$ as I have considered it above have a corresponding structure ‘in the head’ invoking some set of properties, or *features* (as they are called in this literature) to describe the objects recorded in semantic memory? The answer is a resounding and very robust yes.

Most relevant of all the evidence connected to semantic memory and its degradation or impairment, in regard to this question, is not evidence that semantic memory can be impaired, but rather that it can be impaired selectively. Numerous studies beginning from Warrington (1975) have described patients for whom only *some* and not all sorts of things are hard or impossible to name or describe using language. Certain patients could effortlessly define abstract adjectives like ‘arbiter’, but no concrete words at all, like ‘acorn’ (Warrington, 1975; Hoffman, 2016). Some found it harder to name only inanimate objects (Warrington & McCarthy, 1983), while others struggled only with animate ones and food (Warrington & Shallice, 1975). Other patients have only had difficulty identifying fruit and vegetables (Hart, Sloan Berndt and Caramazza, 1985) and people (Miceli, Capasso, Daniele, Esposito, Magarelli & Tomauilo, 2000), suggesting overall that these “category specific deficits are just that – the result of damage to a category-based knowledge domain” (Miceli et al, 2000, p. 491). In other words, they are properties of entities being independently stored and so independently disabled: the building blocks for description.

As Mahon & Caramazza (2009) observe these effects are very often independent of modality: whether selecting the object, naming the object, answering questions about the object, these impairments could (though there are exceptions to this rule, e.g. Druks & Shallice, 2000) transcend exact modality or use. As a result, there is solid ground to claim that categories, of the sort represented in logic by predicates and so able in principle to form descriptions like $Fx \wedge Gx$, are part of human cognition. Whatever else is the case for conceptual cognition it does appear that a descriptive classification element is part of it.

ii. Connections

The second stop on this very selective tour of relevant parts of human conceptual cognition concerns the other vital aspect of descriptions: linking of properties together to classify entities. For this, I will divide my emphasis, between the neuropsychological picture I have used to motivate classes, and am now going to extend to the links between those classes which make concepts descriptions proper, and the behavioural study of the consequences of these links, namely categorisation and feature inference. These links and their consequences relate strongly to those foibles of descriptivism I have considered in §2iii; such that what were mysteries or inconsistencies for the descriptions of Russell (1905) could instead be considerably easier to understand for the sort of descriptions conceptual cognition provides.

Picking up where the previous section left off, conceptual cognition was shown to involve features, or properties (I will use the two interchangeably here). Something logs an entity as animate or inanimate, a vegetable or fruit, or a person – such that when it is damaged, only the things with that property will be more difficult to identify. Taking this idea further, by combining and cross-referencing a large span of cases like the above from neuropsychology, and also cases from cognitive neuroscience (cf. Mahon & Caramazza, 2000), it becomes possible to not just identify that there are properties of objects stored by the human brain *qua* properties but also map these properties relative to each other. In other words, it becomes possible to track the connections between properties that make some of them superordinate to others or that link two particular properties more closely together than a third. These connections of lower-level properties to some higher-level superordinate property, bringing the lower-level properties themselves closer together, form the clusters in the map typically known as concepts and/or categories (I will again use the two interchangeably here) by psychologists at large (e.g. Murphy, 2002). So there are features or properties, like having fur or having four legs, and superordinate concepts or categories that connect these features together to create a reliable box in which to put things satisfying enough of these properties, and from which to exclude things satisfying too few. This is the job of *categorisation* in the sense that psychological science studies it. Features (low-level classes) together form categories (higher-level classes) and every object encountered, pictured or named must be allocated membership.

It should be clear from the way I have just described the traditional definition of concepts that there is (or there seems to be) what amounts to a nested description structure in play. Properties at level N can together form a collection of elements that somehow defines a superordinate class at level N+1 – like having fur and four legs are for being a cat, rather than a canary. And then that class itself may form a collection with other classes of its level, defining the next class higher up the chain – like cats, jaguars and tigers are all felids; and like felids, canids and primates are all mammals; and so on… Even prima facie there appears to therefore be not just descriptions but a schema of *descriptions of descriptions* in semantic memory. Understanding how this schema is curated and deployed presents complexities that together occupy the study of categorisation, much of whose tools are formal or mathematical (see e.g. Pothos & Wills, 2011; Danks, 2014), and which I only intend to discuss in the context of one solution.

What is however clear across many of these approaches is that there really is a statistical basis (as was anticipated by the later descriptivists) to how conceptual categorisation is resolved. Certain properties, viewed by participants as more ‘central’ or ‘(proto)typical’ (see e.g. Medin & Shoben, 1988; Sloman, Love & Ahn, 2008; Jee & Wiley, 2014), are weighted more heavily for categorisation decisions than others; and different properties are prototypical to different categories. For example, Medin & Shoben (1988) observe that participants were not willing to categorise a non-curved shape as a boomerang but are willing to categorise a non-curved shape as a banana. As a result, assuming some means for giving a conceptual version of descriptivism (which I consider as a whole in the next section), the problem of inconsistency, where mistakes can be ignored when otherwise classifying something as e.g. ‘Aristotle’ would appear to be resolved by, precisely, ignoring some of those properties based on their typicality. That is to say: there is a systematic way to ‘grade’ properties and so weight classification decisions in accordance to the most important of those properties for each concept. Setting aside the finer detail of how this weighting works, there is no real debate against this phenomenon. If concepts were to offer a more cognitive sort of descriptivism for information content then inconsistency is not a problem for it. Rather, it would be part of the definition of such a theory that conceptual descriptions are statistical in nature and it *is* part of the definition for the particular theory of conceptual description I will consider.

Returning to the building blocks of any such plausible theory. there are countless studies that support the more basic facts of conceptual organisation, that feature-level classes are related to each other and that concept-level classes supersede them. In neuropsychology alone, there is evidence from the same group of studies I have already discussed for e.g. impaired naming of just subordinate properties such as being made of metal, or being dangerous, as opposed to superordinate categories, like being a bird (Warrington, 1975) and of impairments at any one of three levels (often called ‘features’, ‘concepts’ and ‘classes’, though I use ‘classes’ more broadly here) while all the others are all spared (Crutch & Warrington, 2008). Combined with work in neuroimaging there is enough evidence overall of those features connected together within and across distinct modalities like shape, action, motion, texture, colours and verbal labels to enable the prospect of *mapping* intra- and inter-category relations – both relative to each other, and relative to the topology of the brain (Lambon Ralph, Jefferies, Patterson & Rogers; Patterson, Nestor & Rogers, 2007). In the more extended behavioural literature, seminal work from Rosch (1975) set the stage by identifying that a superordinate class label, like ‘lion’, led to faster identification for labels of properties subordinate to a *related* class, like ‘stripes’ (related to ‘tiger’ that is itself related to ‘lion’ by being subordinate to ‘feline’ – for an even longer chain see also McNamara & Altarriba, 1988). This semantic ‘priming’ effect – where priming is understood in psychology as the facilitation (plainly: the speeding up) of some later instance of cognitive processing by a previous and somehow-related instance (in this case reading the word ‘lion’), has been followed up by other effects of priming both within and across modalities – the linking of categories from language with effects of perception and action (e.g. Zwaan & Taylor, 2006; Glenberg & Kaschak, 2002) likely being the most famous in having motivated the view that conceptual classes can expand classification information by adopting information from across modalities (e.g. Hoenig, Sim, Bochev, Herrnberger & Kiefer, 2008).

Beyond modalities, it has also been argued in the recent literature that conceptual categories adopt not just information from different modalities, as part of their inherent content, but also information about context. In other words, that concepts inherently include information about the context where they are being applied whether over long-term or very fine-grained timescales (Yee & Thompson-Schill, 2016) and as a result they change to fit the place and time. Casasanto and Lupyan (2015) capture this idea in positing that concepts are uniformly ‘ad hoc’: that there is no context-independent ‘dog’ or ‘cat’ class, because there are no features of ‘dog’ or ‘cat’ whose relationship to the classes context cannot modify. If this is correct and concepts can be accepted as a style of description in the Russellian sense then the non-modal problem of context for descriptions is thus solved: not by changing the ‘job’ of description from information content to referent selection and back, but by redefining the properties each of these descriptions appeal to, based on the context for which a referent is being described. For the busy room imagined by Donnellan (1966) it is thus acceptable that ‘martini’ *is* ‘water’ when taken as concepts if they make interchangeable categorisation decisions for that context. For another context they can still be different, or rather: the counterpart categories associated with the labels for that context may differ. This flexible view of properties can appear alien or counterintuitive, but it is as functional as the more immutable one associated with ‘discovering’ (or as I consider in the next chapter *baptising*) persistent natural kinds. If this work is right, it just replaces one ‘analytical fiction’ (Barsalou, 1987) for another.

I return to some consequences for the apparent span of information that concepts appear to encompass later below. For now, a more relevant question for the remaining concerns raised earlier in the chapter about descriptivism (whose original form I am arguing a conceptual theory may replace) is the upshot of those connections behind semantic priming and the selective impairment of a set of entities sharing a category. This last aspect of conceptual cognition I consider in this section is the routine human use of its complex network of hierarchical interconnections for *feature inference*. Feature inference (Rips, 1975), also called category-based induction (Osherson, Smith, Wilkie, Lopez & Shafir, 1990) is using knowledge of some part of the set of categories and features an entity falls under to infer its remaining features. This can take a (more or less) direct form: for example, every mammal is warm-blooded so a new mammal can be safely inferred to be warm-blooded *because it is a mammal*. Other inferences are far less clear: if alligator embryos have no sex chromosomes, would crocodile embryos not have them as well (Fisher, Godwin, Matlen & Unger, 2015)? If not, is it because they are reptiles (inference from known category), or instead because they have a similar appearance (inference from known features)?

As with all of the above, there is a vast literature on what guides the outcome of feature inferences, in both adults and children. One strand of that work explores the contrast between category membership and known features (e.g. Murphy & Ross, 2010; Gelman & Markman, 1986). Another strand explores the varied influence of features in inferring other features, based on how typical of their category they are taken to be (Osherson et al, 1990); or the varied influence of categories in inferring features, based on e.g. whether it is perceived as a natural kind concept (e.g. Gelman & Davidson, 2013). The latter in particular implies conceptual categories also encompass meta-level beliefs over concepts (Malt, 1990) such as whether that concept should be more consistent across contexts – and hence more informative. What is important about this literature for present purposes is however the more basic fact that feature inference happens in the first place. Whatever precise interrelations may guide this process the known conceptual categories and features of some entity (or referent) beget the inference of its other features.

Adding this process, whereby $Fx \wedge Gx$ (F and G being either concepts or features) implies a far larger description that includes all of their implicit interrelations, makes the rest of what plagues the original descriptivist theory of content easily resolvable by an updated, ‘conceptual’ descriptivism. If concepts and features (whose individual roles and interrelationships I will consider in the next section) are used as the building blocks (the *Fs* and *Gs*) of some description they would not come alone to the party: for each explicit category or feature there will be many more inferred of whatever is known to be *F* and *G* to start with, such that the real descriptive specification of this referent would far exceed these explicit elements. Another way of putting this would be to say that for a conceptual descriptivism there are no explicit descriptions at all. Instead there are determinate implicit descriptions based on inference from whatever properties (features or categories) are explicitly given, and *those* are the information content. Clearly, for this and all of the above *specifying* how these processes are (as I put it above) curated and deployed – how concepts are acquired, how categories are selected, etc. – is an ongoing matter for the empirical approach to how exactly human cognitive agents categorise and infer features. A productive link can nonetheless be made between parts of that behaviour and a descriptive account of information content: there are properties, there are categories, there is a statistical aspect, there is feature inference.

iii. Concepts as Hypotheses

Above, I have considered prominent aspects of feature, concept and categorisation behaviour, studied by empirical psychology, and their links to problems faced by the descriptive approach to information content. In doing this I have taken a loose view of the subject matter. I have not attempted to limit this discussion of concepts to concepts connected to words – and I have also omitted a discussion of labels *qua* labels, the lexical entities somehow involved in this process, which e.g. Russell (1905) eliminates. More importantly, I have not considered any one theory of concepts but rather the more general trends that appear to underlie this sort of behaviour. In this section I discuss one particular theory of concepts in whose terms it should be possible to give a theory of information content for reference (as has been my ultimate goal for this chapter), developing the idea that concepts are themselves theories (Gopnik, 1988): notably theories expressible as statistical hypotheses (Gopnik & Wellman, 2012; Danks, 2014).

The roots of this approach to concepts, at least in its modern and computationally elaborated form, are in the modelling of causation – and the development of a certain kind of computational model that can capture its structure (e.g. Pearl, 1988; 2009). What is most problematic about causation, as understood in both the computational and the philosophical domain (e.g. Lewis, 1973), is that associations are not necessarily causes. Two events might co-occur very regularly without one being the cause of the other and the standard way of thinking about what makes something a *cause* of something else, inherited by Lewis (1973) in particular, is to think of counterfactual versions of the same system: an alternative set of facts that has often never happened, where something is different while everything else is the same. For example, in reasoning about whether smoking cigarettes causes cancer it is not enough to know if smokers get cancer. What is instead required is some sort of what-if scenario, where that same person who is a smoker chose not to smoke but did everything else the same; what is required is some means for reaching out, from the actual turn of events we are trying to explain, and into the hypothetical turn of events where a suspected cause is removed to test its influence. Of course, often enough this might be approximated by a physical sort of model, i.e. some other person living an identical lifestyle as the smoker who got cancer, and who is not a smoker. A comparison could be drawn between (enough of) them per basic null-hypothesis driven scientific inference (e.g. Cohen, 1994) to determine if there is a significant difference in cancer onset between two groups of similar-lifestyle people, bar the smoking.

Nonetheless, an obvious intuitive response to even the standard statistical solution, well-trod as that is, would be that something else may be affecting the outcome. If all the smokers in the study were made to smoke outdoors as a result, perhaps this doubled their exposure to airborne pollutants. Was it really the smoking that caused them to have more cancer, or the airborne pollutants? Perhaps a little of both? Once again the only way to grapple with this question is to alter only one element, etc., which is soon well past feasibility to achieve. More than that, some elements cannot be manipulated easily and some causal relationships (as with the case of ‘a little of both’) are more complicated. Most important of all, for present purposes, the question of how human cognitive agents come to understand causal relations – the way that we work out causal dependencies – cannot rely on manipulating the world alone. There must be some way of ‘reaching out’ to systematically imagine what is not the case, set against what is.

What has offered a solution to these problems is the computational structure of ‘generative’ models in general (cf. e.g. Ng & Jordan, 2001; Bishop, 2006) and *graphical* generative models in specific (Pearl, 2009; Koller & Friedman, 2009). The generative class of statistical models is distinct from the kind of so-called discriminative model the hypothetical cancer study I have just described would typically use in that it does not model an outcome (cancer) based on the input (smoking) but also the input based on an outcome. To make this clearer it will help to consider a very basic example of this species of model from Pearl (2009), where a man in a casino declares “twelve!”, and we ask *what game he was playing*.

Even knowing the exact variety of games available, namely roulette and dice, and knowing how many instances of each game are running concurrently, this is a challenging question to solve by comparing the dice-rollers and roulette-players to decide which of them get the number 12 more often as a whole. But it is possible to produce the answer much easier with the same information, by 'reaching out' past the input of "twelve!" and into the entities producing the input using (elementary) Bayesian inference. Using Bayes' Law, the probability of *each game given the outcome* "twelve!" – or P (A | B) – may be given as a function of the probability of *the outcome given each game* – or P (B | A) – conditioned by how many of each game are ongoing at the same time – or P(A) and P(B). What that amounts to is an inversion of the problem, in a way assuming an answer, then estimating the likelihood of the question. For the case of human cognition, this sort of inverted structure has been useful for modelling how that sort of thinking behind what-if judgements, especially causal ones, can be made effective and efficient by starting with an answer and fitting it to a question rather than the other way around (Sloman, 2005) – and many mathematical tools and approaches from that literature were then carried over to concepts, including the basic information unit of beliefs along a gradient being attached to alternative outcomes.

The other part of this 'causal revolution' in addition to generative modelling overall, was the idea that information can be represented graphically in a way that constrains computation (Pearl, 2009). Such a 'graphical model' encodes these same beliefs along a gradient, e.g. I expect it will rain today but I am fairly certain there will be no snow, but with an additional element representing dependence relations: that is to say, whether the values of two arbitrary nodes representing e.g. my belief it will rain and my belief that it will not snow, are computationally dependent on each other. Combined with the 'Markov assumption' (Pearl, 2009; Danks, 2014) that nodes are computationally *in*dependent except when they are connected, the stage is set for a way to track not just causes and consequences, but also the precise extent of their reach. In other words, that point at which something is no longer relevant, illustrated by the macabre (but memorable) example of a firing squad in Pearl (2009), as replicated here in **figure 2**.

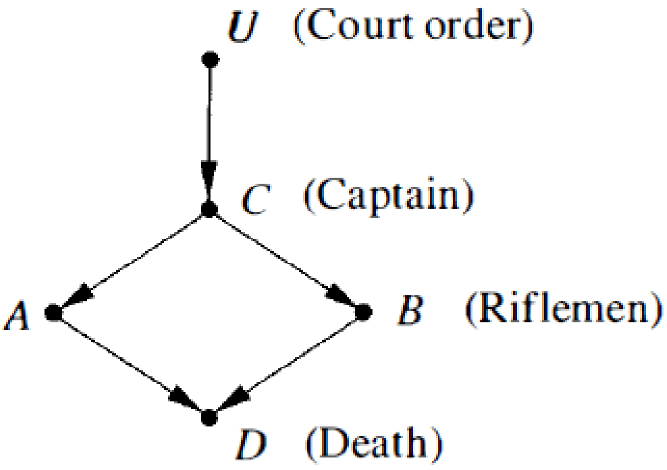


Figure 2. Death by firing squad. From Pearl (2009, p. 207).

In this scenario, what Pearl (2009) is trying to illustrate is the fact that Death may be caused by more than one node of the graph: if Rifleman A believes they must shoot or if Rifleman B believes they do. Yet due to the directed acyclical nature of this graph (the finer details of which are not important) that node representing (belief that one faces[14]) Death need only be triggered by one of them. If Rifleman A pulls the trigger, it is no longer necessary to consider what Rifleman B might be doing. Further afield,

[14] Pearl (2009) does not present this specific graph as representing belief but my modification is inconsequential.

as long as information on either rifleman is available there is no value in knowing whether the captain has ordered the execution: his part of the causal chain has become unnecessary to examine. As a result of this sort of recorded dependency, just like with the inversion of the problem, it is possible to model a complex system in a way that only really depends on 'local neighbourhood' relationships to resolve.

Returning to concepts, the success of causal models of cognition was expanded into a previous theory of conceptual cognition known as theory-theory (Gopnik, 1988) where each category is understood as a theory of how features are combined, tested by observation, and then revised where necessary. With the new computational tools that I have just briefly discussed, the theory-theory became elaborated as a form of graphical model theory (Gopnik & Wellman, 2012; Xu, Dewar & Perfors, 2009). And more recently Danks (2014) has demonstrated that even the other main approaches to conceptual cognition, casting concepts as idealised collections of representative features known as 'prototypes' (e.g. Posner & Keele, 1968), or instead collected features from an example set of previously observed members of a category, known as the set of 'exemplars' (e.g. Medin & Schaffer, 1978), may be given as graphical models. At any rate I will not be pursuing the contrast of individual theories, or their internal nuances, further. Whether or not it is right to say the graphical models approach can 'unify the mind' as Danks (2014) claims – or unify theories of conceptual representation and learning – I will focus on it hereon.

According to the form of this theory presented in Danks (2014), concepts can be gainfully understood as probabilistic hypotheses (in his terms, distributions) presenting the range of options for each feature of a concept, such that each concept (like 'cat') has several possible features as its dependents some of whom are more or less likely of cats. This captures the 'weighting' element of categorisation but what is more important for my purposes is how graphical models track sub- and superordinate relations and context. Specifically, a graphical model of a concept like the ones in Danks (2014) may be understood as a collection of likelihoods for (effectively) subordinate features, like having four legs or long fur or similar, themselves linked to derived features further 'down the line'. The links between features, and the links between features and other elements representable in graphical models like cause-and-effect, are one of the most valuable aspects of the graphical model approach in specifying to a computational standard exactly what 'goes into' concepts, and equally what 'goes out of' them, to inform actions etc.

More than that, the ability of node dependencies to 'toggle' the flow of relevant information on or off also gives an automatic representation of context, as Danks (2014) discusses in the context of placing different concepts with connected features in the same graph with a toggle between them per **figure 3**.

In this case switching between something being a cat or dog entails simply switching what is relevant to the current information context via the topmost node. When there is a cat, only 'cat' dependencies are active. And both the feature nodes and the switch itself are described in the same sort of language and as the same sort of entity. Accordingly, context of use can be accounted for in a very similar way. And another valuable aspect of these toggles is the capacity of this framework to create hierarchies by stacking these graphical models on top of each other; in the sense that what was once superordinate is now the 'ground level' to a higher order of superordinate concepts. As Danks (2014) suggests, dog or cat can become 'features' (subordinate properties) in the category of 'mammal', then 'mammal to the category of 'animal', etc. Lastly, this hierarchical structure is not limited to conceptual cognition. It is

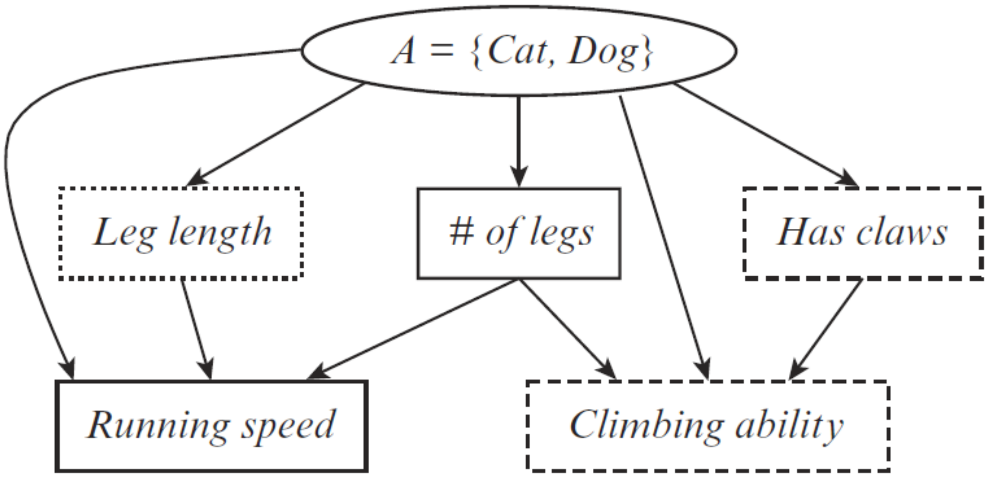


Figure 3. Cat or dog ‘context’ toggle. From Danks (2014, p. 117).

also popular in models of perception, whose homomorphism with the above will be critical in the next chapter as I further develop the link between content and contact for perception – and so for reference.

And as given here this sort of graphical theory form for concepts affords one last exceptionally useful feature. The graphical models in this literature are generative, expressing probability distributions for concepts based on what features they each *ought* to have, conditioned by prior observations (be it via exemplars or prototypes). They operate much like the example of “twelve!” by trying to capture how likely it is that something which is a ‘cat’ would have certain features, rather than how likely each of the observed features is to belong to a ‘cat’: like the case of “twelve!” they model the source of input. That is, they are not modelling cat-appearances, or cat-features, they are modelling hypothetical cats; and they compare subordinate features against this model of a cat, e.g. to determine if a cat which has long whiskers ought to also have long fur to resolve feature inferences, based on that hypothetical cat. (Again, be it an exemplar or prototype: I consciously gloss over the exact form of the construct here.) And being designed to model hypothetical cats against input, to compare with e.g. hypothetical dogs, that also means that these graphical models may produce a hypothetical cat with no input given at all.

Because graphical models are a language, that even reduces to first-order logic under some conditions (cf. Kimmig, Mihalkova & Getoor, 2015), the ‘hypothetical cat’ I have just discussed will be given in the form of a systematic description in that language. It will describe the cat, not relative to sense-data but relative to features, themselves possibly ‘hypothetical’ if they are originally derived the same way, encoded in nodes of the graphical model, so that for the purpose of that model in its current state *that is a cat*. And because these models incorporate at least a number of all the influences not traditionally captured by conceptual theories in the same frame like the effects of action, and the effects of context, then then there is no step required to translate these into a description as they are already a part of one. What is mainly left at this point is just to account for the role of labels qua labels as part of this theory. Ought their labels be features of concepts? Ought they be represented in these graphical models at all?

iv. Labels as Cues

I conclude this discussion, and the chapter as a whole, by considering the role of labels qua labels for conceptual cognition. Russell (1905) was explicit in denying labels a predicate role, except as proxies for description; others in the philosophical literature argue that labels should be included as predicates (e.g. Burge, 1973) in descriptions. Kripke (1980) contends, as I will discuss in the next chapter, that a label, or in his terms a name, performs a wholly different role than description and is irrelevant to one. The side I will ultimately take on this issue will most resemble the second view: that labels, qua labels rather than as stand-ins for concepts, themselves form a part of the information content concepts store. At the same time however, this part is closer to a proxy of the concept than a feature, and I will spend this section explaining the reasoning for considering concept labels in this manner, and for collapsing every referring expression involved in object reference – proper names and common nouns – together.

Lexical labels in general are studied by cognitive psychologists as part of a monolithic mental lexicon, or as Bonin (2004) puts it, "some words to talk about words". This lexicon is the assumed store of any lexical strings and is understood to contain the literal list of known strings of letters – or lemmas – for usage in language production (see e.g. Levelt, Roelofs & Meyer, 1999), and language comprehension. What this lexicon can do in a conceptual capacity is however a question only recently approached, by Lupyan (2015; 2012; 2008; Lupyan, Rakison & McClelland, 2007) and others (Malt & Sloman, 2007; Xu, 2002) with emphasis on the role of linguistic (category) labels in facilitating cognitive processing. The overall spirit of this modern literature is captured by Lupyan & Lewis (2017) as a transition from understanding lexical labels as simple mappings onto a concept (which is also roughly the view based on philosophical descriptivism) to an understanding of lexical labels as a set of cues for categorisation and conceptual inferences. In this new guise, labels are seen as one source of input among others, like perception and action, directly involved in constructing concepts, rather than just 'marking their spot'.

Category learning in particular has a longer history than any other in treating lexical labels as a source of robustness to newly-learned concepts. Xu (2002) describes them as 'essence placeholders' for later learning in children, and work like Sloutsky (2010) suggests that labels are part of the maturation of a perceptual category into a fully-fledged conceptual one. Lupyan et al. (2007) report that the presence of a symbol (cf. Clark, 1992) alone made categories easier to learn even for adults, although Lupyan (2008) reports this has also made them harder to remember – implying that adding a concept label to some novel category is not necessarily as simple as adding a new feature. It might instead, as Lupyan (2008) argues, shift the representation of that category into another format. Along such puzzling lines is the later observation by Lupyan (2015) that use of labels can elicit prototype representations, where the same task otherwise elicits exemplars. Such overlapping use of different categorisation paradigms is not in itself new (see e.g. Danks, 2014; Weiskopf, 2009; Ashby & Maddox, 1993) but had not been linked to labels. Labels thus seem to tap, and even change others into, specific types of representation. The relations of lexical labels with non-conceptual cognition is also noteworthy, boosting objects into visual awareness (Lupyan & Ward, 2013) and outperforming purely visual cues for the same category (Lupyan & Thompson-Schill, 2012). Lupyan (2012) theorises that labels can function as feedback for visual object perception giving direction to the relevant inferences (as I consider for the next chapter).

Most important of these ideas for present purposes is the bigger picture view of 'language-augmented thought' (Lupyan, 2012) or the 'centrality of language in human cognition' (Lupyan, 2015) or of how 'language programs the mind' (Lupyan & Bergen, 2016) where the implication is that category labels, and by extension purely linguistic input for conceptual processing, may create and manipulate certain types of representation by recruiting the sensorimotor-oriented information content from categories in the service of education, communication, and outright imagination. Lexical labels can also furnish the members of sociolinguistic communities by the process alignment (that I consider in the sixth chapter) with conceptual categories they would otherwise lack – such as blind people given concepts of colour. This last aspect of labels is what most motivates my assumption that, were the graphical models from the previous section extended to include labels in this augmenting capacity, the main role for labels in that (extended) system would be best compared to features helping to construct and revise hypotheses. Nonetheless, the representation-shifting aspect of labels implies their presence would remain special – at least in their weighting against other features – to account for their great hypothesis-guiding power.

It is lastly this great hypothesis-guiding power that motivates me most of all to think of proper names and common nouns, so far as they are both 'thing concepts' (cf. Sloutsky, 2010), as two tokens of the same type: lexically-cued categories, one of which has an overwhelmingly stronger cue than the other. This mirrors the idea introduced earlier (in chapter 3) that identifying a proper name is easier because their labels are predicates satisfied by just one object; now expressed via concept identification being more facilitated by labels than other features, and unique-label concepts being identified fastest of all.

4 Augmented Description

In this chapter, I have aimed to cross the span between two separate worlds in the study of reference. To begin with, I took the theory of descriptions by Russell (1905) as the par excellence philosophical account of information content for reference, explored its reasoning (that I will return to in chapter 5) and its weaknesses. My emphasis here was not as much on its core weakness with regard to the modal objection offered by Kripke (1980) for contact, but rather the weakness and unclarity over parts of its role in specifying precisely what information content is attached to a label, and extended to a referent.

I then considered the structural similarities between descriptivism and conceptual categorisation, and the means by which conceptual cognition in general, as studied by psychological science, can address these complaints as part of its more extended set of mechanisms like inference and context-sensitivity. Through its dynamic, flexible representations that 'grow' additional information by virtue of a hidden network of connections between features and concepts, I have tried to demonstrate that descriptivism can be revisited as a theory of conceptual information content based on hypothesised entities and their properties, in the form of generative graphical models. Finally, I have considered the role of labels for this view of conceptual cognition (and by extension the information content of reference) arguing that the best place for labels is as powerful feature-like cues involved in constructing and testing concepts.

In all it is possible at this point to revisit C2, the definition of content I offered in the previous chapter:

(C2) CONTENT

For some language $\mathscr{L}$ and set of labels L, I take the *content* of L to be the specification of all L in $\mathscr{L}$.

From what I have aimed to argue, the language of a generative graphical model describing conceptual categories is a strong and plausible candidate for describing the information content linked to labels in reference; one that encompasses inference, context and other appropriate elements of this information.

So long as it functions the same as a context node, even the existence of referents as opposed to being 'empty' can be represented as part of this language: as it must have a label represented in the model, a label with no referent will not be empty, in the sense of not existing at all. It will instead be part of the graphical model as a 'vacant name', where the hypothesised label-bearer is well-defined yet is just not available for the present context. Assuming historical contexts can be represented like counterfactuals (as they inherently are), empty vs. non-empty could then be toggled across them by context nodes. At the same time this does not address the class of 'mistaken names', whose presence in the list of labels is effectively an error like phlogiston. And together with the modal objection, these still argue against description even in this augmented form as sole determiner of contact; which I move to consider next.

Chapter 4: Contact

1 Building Bridges

> "I hope you may agree with me now that in representative knowledge there is no special inner mystery, but only an outer chain of physical or mental intermediaries connecting thought and thing. To know an object is here to lead to it through a context which the world supplies." (William James, 'The knowing of things together', 1895, p. 109.)

I now turn to contact, as the bridge between labels, information content and referents. In this chapter I will be defending the thesis that not only is such a bridge possible, but that it is entirely commonplace and has widespread acceptance in a different domain, namely object perception. I argue this domain is not as different as it may appear, and that an account of object reference can be crafted from the same building blocks as an account of object perception, applying the same approach for the same concerns.

Key to this will be the two-step mechanism of description and revision that explicitly underlies object perception and that I will argue extends to object reference. Such a two-step process permits progress from our internal concepts to an external world, in a way that fulfils the role of a Kripkean account of contact, by invoking a descriptive theory of content as equal part of its success. The result is a system where reference reaches out to objects in the world but whose means of doing so not only *can* remain grounded in the internal machinery of description, but it *must* be grounded in description to allow this. Equally key will be a presumption underlying this machinery that its output is accurate: that it is right.

Yet much needs to be done to motivate such a story. I first of all consider the processes behind object perception, generally exploring the idea by von Helmholtz (1867)[15] that perception is fundamentally – though with different significance for different traditions – mediated by inference. Throughout this I highlight two fundamental features of visual object perception: a) its underspecification by input and b) its specification of output, in spite of (and sometimes following) error. Taken together I will argue that those properties allow perception to turn very negative odds into very positive odds in its favour.

Over the latter part of the chapter I will then consider how the mechanism of perception may relate to the mechanism of contact in the context of reference, in the context of the final major theory from the 'classical semantics' trio of philosophers. As with the last two, the view by Kripke (1980) that contact is a function of causal linkage is not one that I will accept as given, but rather explore its usefulness – and its weaknesses – with the aim of constructing or appealing to a system that may fulfil such a role. In the final section, I do just that – by appealing to perceptual *representation* as a criterion for contact. This 'statistical' theory of contact then serves to inform the framework I motivate in the next chapter.

[15] Potentially itself descended from as early as Ibn al-Haytham/Alhazen (1083/1989); see e.g. Howard (1996).

2 Objects from Light

i. Inference

At its most minimal, human visual perception is the transformation of light signals through electricity. In any given moment, the human retina is bombarded by light that is either emitted into it or reflected into it by objects in the world – the ambient optic array (Gibson, 1950). This heterogeneous bundle of light is projected onto the retina as a two-dimensional pattern: the *proximal* stimulus. The pattern then triggers electrical signals that ultimately[16] stimulate different cells of the visual cortex according to the original features of the two-dimensional projection (e.g. Palmer, 1999). These cells then activate other cells, engineering our familiar sense of seeing the source of the original projection, the *distal* stimulus.

If the distal stimulus is a cat the proximal stimulus is a two-dimensional painting of the cat using light. Even if there was no subsequent breaking up of this painting into its features, or any sort of translation of the painting into a series of electrical signals, there is already a brute mathematical fact in evidence. Cats are three-dimensional entities. Retinal projections of cats are two-dimensional entities. It follows that there is even in this extremely simplified case an element of *transformation* from the proximal to the distal stimulus, some sort of 'figuring out' of how the cat must be in three dimensions. This is the province of vision, in the contemporary sense of the word: "the process of discovering from [retinal] images what is present in the world, and where it is" (Marr, 1982, p. 1). I will follow Palmer (1999) – who in turn himself follows Rock (1983) – in marking a further transition from vision to *perception,* as *the acquisition of knowledge about the distal objects* beyond extrapolating their shape; though this distinction will only be loosely adhered to until a more exact definition of 'knowledge' follows later. For now, the real point of discussing definitions is that even vision simpliciter, as an extrapolation of complex three-dimensional shapes from two-dimensional retinal projections, is a task whose success inherently involves a decision over which one of several possible interpretations of a retinal image is the bona fide object reflecting light. Proximal stimuli lack the information to wholly depict the distal.

This principle is easy to see in a very simple case. Take the lines from **figure 1**, whose projections on the arc standing in for the retina (on the left) are all identical, despite differing orientations and length.

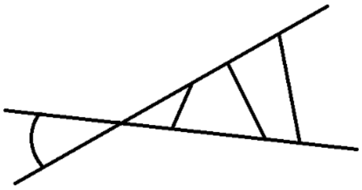


Figure 1. *An inverse problem.* Adapted from Palmer (1999, p. 23).

Knowing only this projection it is mathematically indeterminate which line on the right projected it – one so-called inverse problem (Palmer, 1999; Pizlo, 2001). Only by comparison to retinal projections

[16] An enormous amount of detail is consciously omitted here, from the varied types of photoreceptor cells in the retina (Schnapf, Kraft & Baylor, 1987) to the parallel pathways retinal signals propagate across the cortex from. Nonetheless this abstraction suffices to convey the intended point, that perception is the result of transformation.

by the same lines from different angles can this inverse problem be tackled, when projections of the lines made at different times from different perspectives are combined. One projection may make it look like the lines are identical but others will not, and by combining evidence from multiple retinal projections the possible interpretations of these projections *in toto* will be constrained. According to Gibson (1979) this compilation of *successive retinal projections* from different ambient optic arrays suffices to reconstruct the distal stimuli reflecting or emitting light, so that perception is exclusively and directly linked to proximal stimuli, which collectively encode all the physical facts of the distal.

The *direct perception* Gibson (1950, 1961, 1979) envisions is essentially detective work. Using the ambient optic arrays as clues, facts about individual objects emerge as observers move and interact with their environment. Key to this sleuthing is the notion of invariant and variant elements of such ambient optic arrays: those patterns that change and do not change over time as the observer moves. In a crude version, this is demonstrated by the hypothetical proximal stimulus presented in **figure 2**.

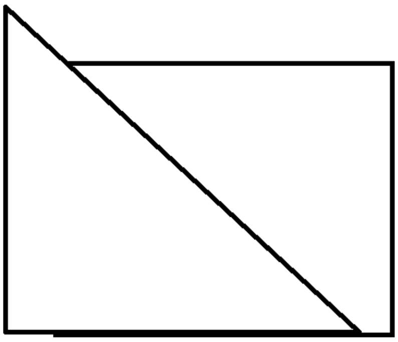


Figure 2. *Perceived occlusion.*

Prima facie this two-dimensional retinal imprint poses another inverse problem for the visual system. For a human observer, this is intuitively a triangle in front of a rectangle against a white background; such that in three-dimensional space, this is likely a pyramidal object occluding a rectangular cuboid. But based only on this image, the distal objects could be counterintuitive things as shown in **figure 3**.

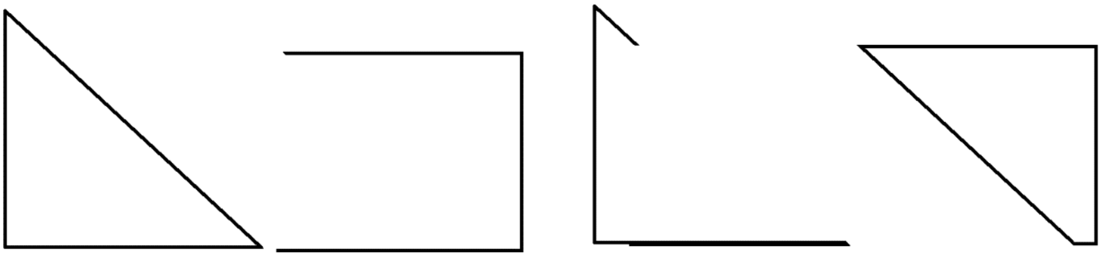


Figure 3. *Counterintuitive cases.*

And even before such cases there is the question of where shapes begin at all in a real retinal imprint. Where does the background of a visual scene end and the edges of each object start in the first place?

For Gibson (1979) the answer to both occlusion and edge-detection is, as mentioned above, sleuthing. Moving around a real cuboid and pyramid, the projections of both objects will vary greatly according to the angle they are being observed from. The projection of the floor on the other hand will vary less.

Edges may be apparent by the way the light reflected from the cuboid and pyramid *changes* when the light reflected by their background *stays the same*. For occlusion the clue is in how it can be reversed: as an observer circles an occluded object, its projections systematically change (e.g. the cuboid might eventually occlude the pyramid instead), and then change back as the initial vantage point is revisited. Perspective-variant aspects of proximal stimuli become clues to the structure of perspective-invariant distal objects by the systematic way they vary as they are observed. And this Gibson (1979) argues is a case of direct perception – not because objects are ever perceived in three dimensions as opposed to two-dimensional fingerprints but because the fingerprints *in toto* suffice to determine the distal world.

Much opposition exists to this idea from Ullman (1980), Gregory (1980), Marr (1982), Rock (1983), Palmer (1999) and others. But irrespective of criticism, even if Gibson (1979) is correct about distal stimuli being recoverable from proximal stimuli alone, there remains an extent to which even 'direct perception' (in Gibson's sense) involves the integration of evidence toward some conclusion. Some set of premises $\{x_1, x_2, x_{3,} \ldots\}$ – here the observed features of retinal projections – must collectively determine a conclusion set $\{X_1, X_2, X_{3,} \ldots\}$ for the properties of distal objects that we then perceive. There is some room for a misunderstanding here worth dispelling: although my present rendering of direct perception uses inferentially-laden language ("detective work," "clues") and I use the schema of an argument to frame the theory, I have no intent to recast direct perception as ultimately indirect. My presentation only echoes the idea by Rock (1983) that perception as a process can be understood as a sequence of descriptions, in this case of each distal property X_n based on features $\{x_1, x_2, x_{3,} \ldots\}$ of some proximal stimuli. So if the shape of an object is determined based on how light reflects off it from a combination of perspectives, these reflections would each be premises leading to a conclusion. The claim is perception could be understood as a logical argument, not that it is implemented as such.

Taking this schema into account, the counterpoint to Gibson (1979) becomes straightforward to state. The thesis of *indirect perception* is that premise sets for perceptual conclusions may contain elements that are themselves perceptual conclusions. In other words, that one perception can depend on another (Rock, 1997) so the distal stimuli will be *reconstructed from reconstructions* and not just appearances. Or: that those properties $\{x_1, x_2, x_{3,} \ldots\}$ used to reconstruct distal stimuli are themselves reconstructed from proximal stimuli each based on a premise set $\{\chi_1, \chi_2, \chi_{3,} \ldots\}$. And that moreover, the members of this premise set $\{\chi_1, \chi_2, \chi_{3,} \ldots\}$ might also include *assumptions* used to infer the set $\{x_1, x_2, x_{3,} \ldots\}$; not purely features of the proximal stimulus, or of a succession of stimuli as envisioned by Gibson (1979).

Indirect perception in the sense considered here encompasses several different theoretical schools and empirical findings under the traditional banners of 'inferential' or 'constructivist' approaches to visual perception: a generally joint tradition from von Helmholtz (1867) to the present. But cases in point for the combining of assumptions with retinal images making 'indirect perception' indirect are also found across the works of the turn-of-the-century Berlin school of Gestalt psychology – cf. Ellis (1938). The Gestalt school as a whole is hard to classify vis-à-vis the inferential nature of perception, its members having collectively influenced all of Gibson, Rock, Gregory, Marr and more (Wagemans et al., 2012). Its signature is the 'organisational' approach to visual object perception, motivated by experiments in perceptual grouping, whose tenet is that the processing of any retinal image *as a whole* is a necessary precondition for categorising its individual elements. In other words, seeing the forest to see the trees.

Grouping experiments introduced the idea that processing part of an image – e.g. to find the edge of a shape – may be at least informed or even superseded by the result of processing the image as a whole. This general principle is easily demonstrated using figures 2 and 3 above. Wertheimer (1923) defines Direction as a grouping principle: note how the hypotenuse of the triangle in figure 2 is not perceived as divided like the rightmost example in figure 3. It is perceived as a straight line, even as it intersects the upper border of the rectangle. And for Wertheimer (1923) the decision – in a processing sense – to respect the original line rather than break it up as in figure 3 is motivated by processing the entire line to reinterpret the processing of its individual elements. Again the theme recurs of my using inferential language to describe a mechanical process occurring at a sub-personal level and in this case involving no assumption *external to the retinal image*, making the example more akin to Gibson (1979) in spirit. Overall, these Gestalt grouping experiments have been widely replicated (Wagemans et al., 2012) and the influence of assumptions in the form of grouping principles robustly detected in different stages of processing – e.g. both before (Schulz & Sanocki, 2003) and after (Rock & Brosgole, 1964) objects are successfully arranged in three dimensions. Throughout, all these assumptions are strictly image-based: *internal* to the retinal image, in the sense that they are just a function of its two-dimensional geometry. Gestalt grouping principles might thus fit the mould of 'indirect perception' in only the strictest sense. Nonetheless this remains a case of combining assumptions with a retinal image in order to interpret it.

My intent with this remains as above. Whether direct or indirect there is a level at which perception is encapsulated by a premise-conclusion framework and is thus a transformation rather than observation. 'Transformation' in the sense that arguments transform elements from a premise set into a conclusion: rigorously and systematically, but not such that the latter is found inside the former 'as is'. Even when this transformation only uses retinal images or data from a single retinal image as its input, there is no scientific account of perception where the relationship between proximal and distal stimuli is reduced to pure 1-1 matching from a retinal image to what is perceived. For Gibson (1979) such a matching is from *many* retinal images. And for a pure Gestalt approach the matching is also from many: from one retinal image *combined with assumptions* based on that image that serve to constrain its interpretation.

Lest it seem overly permissive of me to interpret Gibson (1979) through a logical framework and then use that framework to also encompass other approaches, which I do in the *spirit* of Rock (1983), there is also the *letter* of Cutting (1991), who arrives at a very similar picture. Much like I have presented it, Cutting (1991) takes 'direct perception' to be a deductive inference from premises to a conclusion and 'indirect perception' to be an inductive inference, made deductive through the use of hidden premises: assumptions that are not part of a proximal stimulus inserted as premises in our inference to the distal. Much like Cutting (1991) I do not divorce the result of 'direct perception' from its exclusively retinal origins by integrating it in the same explanatory framework as the 'indirect' theorists. If anything my intent is to use a common language to explore what makes it unique. And in further defence of this, I diverge from Cutting (1991) in considering the process Gibson (1979) outlines a deduction only from multiple retinal images as an environment is explored, as opposed to information from a single image.

This foundational aspect of direct perception is made quite explicit by Gibson himself – my emphasis:

> "Consider, for example, the age-old question of how a rectangular surface like a tabletop can be given to sight when presumably all that an eye can see is a large number of forms that are trapezoids and only one form that is rectangular, that one being seen only when the eye is positioned on a line perpendicular to the center of the surface. The question has never been answered, but it can be reformulated to ask, what are the invariants underlying *the transforming perspectives in the array* from the tabletop? What specifies the shape of this rigid surface as projected to *a moving point of observation?* Although the changing angles and proportions of the set of trapezoidal projections are a fact, the unchanging relations among the four angles and the invariant proportions over the set are another fact [and] they uniquely specify the rectangular surface." (Gibson, 1979, p. 74.)

As a result, I take perception to be at minimum *weakly underspecified*. There is no one retinal image sufficient to determine the distal stimuli producing it. At minimum, either additional retinal images or assumptions of some sort must be added to what is actually seen in order for anything to be perceived.

ii. Competition

Considering visual object perception as inference from the features of some proximal stimulus to the properties of a distal stimulus, e.g. inferring the shape of an object from lines it projects on the retina, my intent has been to illustrate what additional premises each theoretical approach invokes to do this. In other words, what must be added to a single retinal image to allow for its features to be interpreted as projections of a three-dimensional distal stimulus, given the infinite interpretations otherwise valid. For the ‘direct perception’ due to Gibson (1979) any additional premises are additional retinal images. For the ‘organisational’ approach, accounting for perceptual grouping phenomena reported by Gestalt theorists like Wertheimer (1923), the additional premises are assumptions about features of the retinal projection *per se*, e.g. a straight line in a retinal image should be interpreted as one continuous border. In each case these assumptions are necessary: without them a distal stimulus becomes uninterpretable. And I have dubbed such assumptions ‘internal’ to the retinal image, to suggest their geometrical basis.

There is however another class of assumption constraining the interpretation of retinal images. This is illustrated straightforwardly using an example also due to Wertheimer (1923) as presented in **figure 4**.

314CM

Figure 4. *Grouping from experience.* Adapted from Wertheimer (1923, p. 86).

Though it deviates somewhat from my earlier examples of inferring object shape from a retinal image (in that this is written language, whose visual processing occupies a custom niche – cf. Rayner, 1998) the above still offers a clear case in point for how prior experience may facilitate perceptual grouping, and more generally of how a retinal image can be interpreted through the use of *external* assumptions. For most readers, this is clearly 314 CM: a measurement of 314 centimetres. There is however no real reason to group ‘314’ and ‘CM’ based exclusively on the geometry of the above image. No characters are closer or further to any other, nor are (e.g.) the ‘C’ and ‘M’ both tilted in the same direction. What instead appears to constrain our inference specifically to ‘314 CM’ out of all the groupings possible is that we are familiar with the abbreviation ‘cm’. And our familiarity with CM allows us t*o recognise* it.

Our perceptual inference from ‘314CM’ to 314 CM therefore seems to depend on the assumption that ‘C’ and ‘M’ must go together here: an assumption external[17] to the retinal image, no matter how many angles are combined or geometrical principles evaluated. This represents a case of indirect perception in so far as it is based on a retinal image combined with an assumption. Even more than that however, this assumption is itself the previous product of perception (of the word ‘CM’): the very same process it functions as input for. What put this grouping assumption in play is *having perceived that grouping.* By allowing a prior perceptual conclusion into the set of premises the current inference may invoke to group ‘314cm’ the outcome is an argument with perception on both sides: perception from perception.

The mere possibility of perception from perception, which Gibson (1979) was at such pains to remove the need for as to exclude assumptions altogether from perceptual processing, inevitably summons old spectres like the Evil Demon from Descartes (1641/1996). If perception depends on perception, then it can be possible for one false conclusion to lead to further false conclusions, until our every conclusion has been contaminated with error, and we are divorced from the world whose objects we tried to reach

[17] Strictly speaking it may depend on the less complex geometry of ‘CM’ vs. ‘314’ in this particular case but e.g. the case of ‘catanddog’, or some other concatenation of visually consistent lexemes, will create the same effect.

– perception will have become hallucination. On the other hand, embracing the loop from perception to perception motivates a mechanism allowing perception to build on itself that modern work in both empirical science (e.g. Hinton, 2007; Friston, 2005) and philosophy (Clark, 2016; Hohwy, 2013) has seized on to base a view of perception departing from the classical picture using Bayesian generative algorithms to cyclically incorporate both past assumptions and current evidence. I discuss this in §2iv.

At the moment, my aim is neither to dispel nor reassert the influence of past perception on the present, but rather to introduce past perception as one factor among many that can inform perceptual outcomes – *pace* Gibson (1979). And in particular, to emphasise how the influence of the past can compete with the influence of present factors to determine the conclusions of perceptual inferences. As was the case for underspecification I shall only consider a weak type of influence for past perception – that it could be an influence, as opposed to past perception always influencing a present outcome, per the breakout Bayesianism just mentioned. Unlike underspecification, however, the weak and strong (*viz.* Bayesian) view of the influence of past experience are easy to reconcile; which will become important a bit later.

Additionally, the processing level *from* which and *to* which past perception exerts an influence varies. For the case of 314 CM the external influence was from the recognition of a lexical string; which as I discussed previously[18] is a high-level mental representation. It was the final product of past perception returning to inform a perception-in-progress. Another very famous example of the same is in **figure 5**.

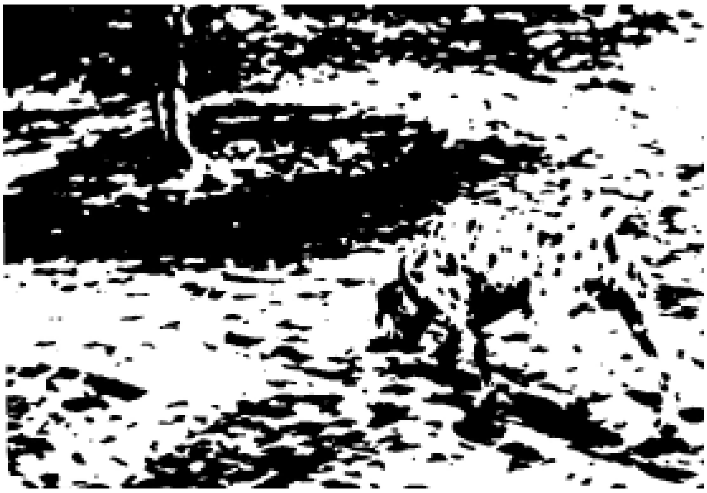


Figure 5. *Past wins over present.* From Miller (1999, p. 650).

For this case the perceptual conclusion based exclusively on the geometry of the present retinal image (which is generally some sort of grouping of dark spots into bigger blotches) is entirely changed when assumptions external to the image inform the process: namely, this is a Dalmatian sniffing the ground. Our *stored past records of perceiving* Dalmatians dictate how this picture is perceived against present factors. But this is not the only case. Rock (1997), whose definition of indirect perception I have used here, also considers the influence from perceptual conclusions of a much lower level: influences from shapes rather than concepts. Specifically, the same influences (from e.g. direction and proximity) that facilitated the perception of the triangle in figure 2, discussed under the guise of internal assumptions.

Though the so-called cognitive penetration of perception by conceptual entities like Dalmatians is not a simple or a one-size-fits-all phenomenon (cf. Pylyshyn, 1999; Vetter & Newen, 2014), the influence

[18] In chapter 3.

of image-based grouping assumptions on the perception of shape is literally plain to see. And as Rock (1997) notes this is also perception from perception, given how grouping itself is a type of perceptual conclusion preceding our perception of the shape to which it applies: e.g. to spot a triangle in figure 2 one has to initially record that the points making up its sides should be grouped, but this is perception. Employing the grouping assumptions inherently demands that some perception has already happened, in the sense of departing from the proximal stimulus. There may thus be two distinct phenomena here:

> "[It] seems correct to say that in both cases, perception is indirect. In the first case – let us call it type A – the indirectness comes about because it is simply not the proximal stimulus per se that can be said to yield the perception we are trying to explain. That stimulus must first give rise to a certain perception, be it of orientation, of proximity, of depth, etc. [which is itself a conclusion about the stimulus], before the final perception can be expected to occur. The second kind of indirectness – call it type B – comes about because some ambiguous stimuli are first perceived on the basis of their literal correspondence to the proximal stimulus. However, that perception or "solution" is inadequate for some reason, so that perception is superseded. (Notice that being *superseded* is not the same as being *replaced* because the prior interpretation still exists.)" (Rock, 1997, p. 14.)

The phenomenon most relevant to my comparison with reference is Type B: that perception might be revisited based on prior perception. It is also worth highlighting the clarification in parentheses about how this preserves the initial conclusion. Thus, in my parlance, there remains an interpretation where figure 5 is analysed using just internal assumptions (the dark blotches), which an interpretation based on external assumptions (the Dalmatian) was selected over. This opens the possibility for conclusions based on *internal* assumptions prevailing instead, under different circumstances, as **figure 6** confirms.

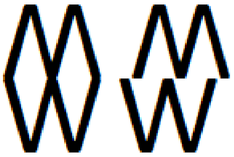


Figure 6. *Present wins over past.* Adapted from Wertheimer (1923, p. 87).

In this further example from as early as Wertheimer (1923) – in deliberate counterpoint to '314cm' – our past perceptual experience of the letters M and W fails to make us perceive an 'M' and 'W' in the leftmost image. A conclusion is instead preferred using just geometrical assumptions: two rhomboids. In the second image where the case for a geometrical conclusion is weakened, an 'M' and 'W' appear.

Competition between present/internal/Type A and past/internal/Type B perceptual inferences drives a stronger underspecification invited by indirect perception: that perception may equally be the product of *some other distal stimulus altogether*. In cases like the Dalmatian from figure 5 the properties used to infer what the image depicts seem equally the properties of figure 5 itself but also *the properties of Dalmatians* (or rather the properties of proximal stimuli caused by past Dalmatian dogs, in contrast to past angles of the same blotchy drawing). Gibson (1979) rejects this example, taking perception to be well-defined only in ecological contexts and not for contrived images intended to fool what otherwise might be an unambiguous process.[19] Ultimately the spots in figure 5 are not a dog. Whether every real Dalmatian dog can be made unambiguous without external assumptions is however an issue of lesser concern, both in the greater literature and for present purposes. What suffices and seems borne out by

[19] "An important fact to be noted about any pictorial display of optical information is that, in contrast with the inexhaustible reservoir of information in an illuminated medium, it cannot be looked at close up [in 3-D space]." (Gibson, 1979, p. 244.)

examples like the previous is that the reinterpretation of retinal images based on external assumptions is *possible* – that prior perception *can* inform and potentially overturn present perceptual conclusions.

An even clearer demonstration of external assumptions, in this case multiple external assumptions all vying to constrain an outcome, is by visually ambiguous images that remain ambiguous like **figure 7**.

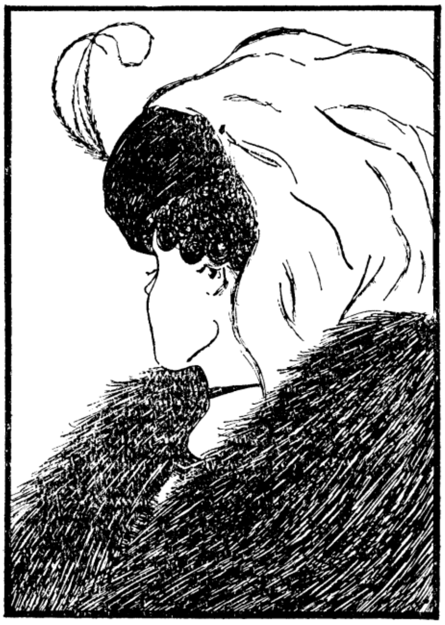


Figure 7. *Inconclusive*. From Boring (1930, p. 444).

This image can be visually interpreted as a young woman turned away from the viewer or as an older woman looking pensively toward the bottom left corner – both wearing a feathered cap of some kind. Each conclusion is driven by external assumptions (records of women or feathered caps) and internal assumptions, such as the contour plausibly marking the young woman's neck or older woman's nose. But neither interpretation can win. Rather than one being selected the two perceptual conclusions just alternate: one conclusion is revised into the other conclusion in an endless loop. The case of 'bistable' images like the above is well-studied and does not appear to depend on conscious awareness (see e.g. Kornmeier & Bach, 2006; 2009): their inconclusiveness is wholly a function of automatic processes.[20]

As a result of these examples – and the processing phenomena underlying them – it is uncontroversial to claim that perception *admits competition* in both outputs and inputs. For every assumption pushing a given perceptual inference toward one conclusion *it is possible* another assumption pushes the same inference over the same proximal stimulus (the same retinal input) toward another conclusion entirely. And for every set of assumptions available from a given retinal input *it is possible* a different set from an entirely different retinal input, caused by an entirely different object, co-determines the conclusion. Lastly, for each accepted perceptual conclusion *it is possible* revision can instead select its competitor. The Bayesian approach to visual perception, which I will revisit near the end of this chapter, becomes simple to summarise in advance: by replacing 'it is possible' with 'it is necessary' in those statements.

[20] Here and throughout I use 'process' just in the most subconscious, mechanical, automatic sense of the term. This is not to say this processing cannot be mediated by conscious attention – cf. section 4 below – e.g. when an observer has not made the connection between figure 7 and the young or older lady and it is pointed out to them: but so long as the competing interpretations *have* both been computed, the inconclusiveness is wholly automatic.

iii. Description/Assertion

Described in the language of inference, perception emerges as a process underspecified from its input and as a result reliant on assumptions to transform any one retinal image into the objects we perceive. It is clear that the source of these assumptions can vary, from rules to other retinal images to previous outputs, and competing assumptions (often from different sources) can support diverging conclusions as a consequence of how perceptual processing is implemented – given examples like all the previous. Such conclusions could then potentially be revised and superseded despite previously being accepted.

In the above, I sought to isolate and highlight fundamental aspects of visual object perception, whose acceptance – in some form or another – cuts across theoretical divisions in contemporary psychology. I now turn to the incorporation of such fundamental aspects into a larger narrative on the mechanism underpinning perception as a whole. In other words, the means by which perception may win against overwhelming initial uncertainty, to reliably decide on accurate descriptions of the objects perceived. An important modern basis for such an overall mechanism is given by the 'structural' (Palmer, 1977), or 'constructive' (Epstein, 1987), or computational theory of perception, founded on neurophysiology as much as the behavioural sort of phenomena I have considered. Taking stock of findings by Barlow (1953), Kuffler (1953) and notably Hubel and Wiesel (1959; 1962) on the fundamental *physical* units that underlie visual perception in the brain, together with how underspecified and assumption-driven perception is in practice, theorists like Palmer (1977), Rock (1980), and especially Marr (1982) came to the idea of a series of decisions – from the certain but underspecified, to the tentative but specified.

At the ground level of this series are the basic physical units, those cells in the visual cortex providing the most immediate facts about the precise pattern of light being emitted into the retina. As Hubel and Wiesel (1959) first showed, some cells will be exclusively sensitive to horizontal lines. Some will just be sensitive to vertical lines. Others will be specifically sensitive to the transition from light to dark or dark to light that often marks an edge for a three-dimensional object; cf. the distinct perceived bottom edge of the rectangle (vs. the triangle) in figure 2. Other cells found by Hubel and Wiesel (1962) only respond to movement, like rotation; or to an object taking up the entire area of the visual field the cell receives stimulation from, as opposed to a smaller portion of that area. Activation within this network of specialised cells comprises the most immediate and basic input available for perceptual processing. As I have discussed at some length now, such basic properties of a retinal image on their own are not sufficient to reconstruct the distal stimulus – the object whose image that is – unless supplemented by assumptions for e.g. putting lines together with edges to produce complete shapes. The computational amalgamation of such basic input into determinate shapes, in order to subsequently decide each three-dimensional object originally projecting light on the retina *based on that amalgamation* – as opposed to the basic input that became amalgamated – distinguishes the hierarchical, computational theory of visual perception most notoriously advanced by Marr and colleagues (Marr & Nishihara, 1978; Marr & Hildreth, 1980; Marr, 1982), but equally anticipated by e.g. Rock (1957; 1980) and Palmer (1977).

In an early manifesto Marr (1974) motivates his computational account of visual perception by appeal to a distinction between two kinds of perceptual processing, which he argues are equally necessary for perception to succeed even in the most basic of tasks. The first kind is *measurement* of a quantity, like light impacting the retina, through its determinate transformation into another, numerical quantity like neuronal activation or inhibition magnitudes. Marr (1974) insists that the output quantity be numerical given his goal to ground visual perception as far as possible in mathematical operations. What is most important about 'measurement' in the sense given here is nonetheless not its form but its basis in cells "acting directly" – as Marr (1982, p. 49) later reiterates – on the retinal image that started this process.

Measurement thus stands for the simplest, causally closest connection between features of a proximal stimulus, and elements of the mechanism used to perceive the distal object that created those features.

Contrasted to measurement, the other kind of perceptual processing Marr (1974) explores is *assertion*. While measurement is a determinate transformation from one input to one (numerical) output – so that each input has exactly one possible output if the measurement is reliable[21] – an assertion expresses the outcome of one or several measurements *in toto* by abstracting from the original units of measurement and often supplementing them with further information. The example Marr (1974) uses is an analogue weight scale where the translation of the pointer's location to some number on the dial it is pointing at is a measurement, e.g. '165', but the overall result *that someone weighs 165lb* will be an assertion that supplements the measurement '165' with further information on the parameters used to derive weight. A correctly-interpreted identical weight scale on Mars reporting '165' would not result in an assertion that someone weighs 165lb – but instead that they weigh ca. 445lb given Martian gravity is 3.711m/s^2.

The summative nature of assertions beyond the apparent, or at least computationally feasible, reach of arithmetical or geometrical operations led Marr (1974) to insist they must be expressed using symbols rather than numbers – and thereby require an explicit cognitive architecture for storing and processing such symbols contra Gibson (1979) and later defenders of equivalent theses like Warren (2005, 2013). The debate around explicit symbolic representation is one I already touched on in the last chapter, and will revisit. Yet it is still worth pausing here to note the distinction of measurement vs. assertion is not limited to how assertions are implemented. Marr (1974) explicitly defines assertions as symbolic. But the thrust of his approach echoes the same sentiment expressed more neutrally by Wertheimer (1923):

> "I might say there were 327 brightnesses and nuances of colour. Do I have "327"? No. I have sky, house, and trees. It is impossible to achieve "327 " as such. And yet even though such droll calculation were possible and implied, say, for the house 120, the trees 90, the sky 117, I should at least have *this* arrangement and division of the total, and not, say, 127 and 100 and 100; or 150 and 177." (Wertheimer, 1923, p. 71, my emphasis.)

Whatever its precise implementation and storage, a *summative step* exists between the basic input that is made available to whatever mechanism underpins perception – like neuronal activation magnitudes caused by different patterns of brightness, quite as Wertheimer (1923) envisioned – and even the least complex output of that mechanism. Something as basic as straight lines that do or do not terminate, or lines that are or are not the edges of a larger shape is still the combined product of measurements from line- and edge- and termination-sensitive neurons; and often not a unique product unless combined, as Marr (1974) initially argues, and Marr and Hildreth (1980) show by considering all the computational outcomes allowed by the same basic input from any single edge- or line- or termination-detecting cell. It is as a result of the need for a summative step that Marr (1974) advocates a symbolic representation, rather than his argument for symbolic representation mandating the existence of summative outcomes. As was the case for concepts I will consequently be accepting and assuming symbolic descriptions of assertion whether or not the symbols do correspond to entities at the level of physical implementation or instead offer a convenient logical description of a mechanism that is otherwise nonrepresentational. If the symbols are just a "descriptive convenience" (Warren, 2005) that will still be entirely sufficient for any appeal I make to their contributions. What matters is the summative step they might represent.

Assertions can decide problems otherwise unsolvable at a single-measurement level by incorporating several separate measurements. Where one measurement might be indeterminate, a system of several

[21] If the transformation of a quantity to a number is e.g. implemented unsuccessfully by the neurons in question, the measurement will be unreliable. But my interest here (as with labels previously) is in how accurate input can be converted to accurate output by some further means; rather than how the initial input itself may be inaccurate.

measurements becomes easier to resolve into a unique solution. But this summative step Marr (1974) demands is not irreplaceable because it can combine individual *measurements* of e.g. bright followed by dark to assert that there is an edge of some three-dimensional object rather than a black-and-white pattern of some other sort like a piano keyboard.[22] Assertion is irreplaceable because even if we might identify lines or edges by summing measurements this is still insufficient to determine a whole object in three dimensions. What is required is a kind of processing dedicated to combining information, not just from measurements of the same kind (e.g. all the edge-detectors) nor just from measurement (e.g. all the edge- and line-detectors together) but from every tenable source, including its own past output. The need for such a step distinct from measurement that can combine information from measurement but also *its own past iterations* is easier to illustrate by revisiting the rectangle and triangle in figure 2. Like single neurons tuned to lines and edges, lines and edges can themselves be combined in different ways. Four lines that are edges can make a rectangle. Three lines that are edges could make a triangle; yet in figure 2, three lines that are edges will still make a rectangle instead. There are multiple reasons why the three lines combine to make a rectangle, as opposed to joining up with a fourth line to make a trapezium or not enclosing any shape at all (as depicted in figure 3). For example, the fact three 'edge' lines straightforwardly combine to make a triangle, leaving just three more available to build a second shape. Or how the back-to-back edges running along the bottom imply two overlapping closed curves.

Yet neither of these reasons or others like them is given simply from lines and edges. The first one, to do with the triangle, requires the assertion that there is another shape. The second reason requires that a constraint is respected – what Wertheimer (1923) would call a grouping principle – regarding edges in a certain configuration, such that they are not combined in ways that allow unnaturally open curves. In every case the determination of a rectangle or a triangle is an assertion about assertions, combining previously-asserted lines and edges, to in turn assert a more complex structure. In the former example, the assertion that lines and edges make a *rectangle* will incorporate the result of the separate assertion about certain lines and edges that form a *triangle*; such that this rectangle is asserted equally based on lower- ('line', 'edge') and higher-order assertions ('triangle'). In the latter case of back-to-back edges the assertion of two closed curves is given equally from measurement and a general rule about curves. And an assertion of a rectangle in figure 2 includes a fourth, occluded line that could not be measured: a line whose assertion cannot be attributed to any measurements, which is nevertheless being asserted.

Demands like these are why Marr (1974) is at such pains to set off by introducing a step distinct from measurement, described and perhaps also represented by symbols, that can combine information with no concern for its type or origin. Measurements, rules and prior assertions can all be combined in this summative step to produce novel assertions whose encoded information (*that there is a line*) becomes the input for more. Based on measurement by a set of activating neurons, assertions are made of lines, edges and the overall presence vs. the absence of a larger object, later termed a 'blob' by Marr (1976), and this is what Marr (1974) declares the purpose of low-level vision: an assertion from measurement. By combining information[23] from one set of assertions to enable another it becomes possible to decide specific features of the retinal input, like the number of objects in an image, their exact shape, size and texture, and location in three-dimensional space. I use the term 'decide' here – as e.g. the rectangle in figure 2 was not purely assembled from the lines and edges present in the image: recall the fourth line that was asserted as part of the rectangle, rather than measured from retinal stimulation. Measurement

[22] To be clear: the relevant alternative to 'there is an edge' is not that 'there is a piano keyboard', but instead that the pattern of brightness and darkness triggering edge-detecting cells in the primary visual cortex is *not* an edge. It is informative to remember at this point that Marr (1974) originally construed assertions as logical statements.

[23] This is often quite literal, such that e.g. the convolution of different input functions into a new one is achieved via integration in the mathematical sense (cf. Marr, 1982), but my intent with the term is more broad than $y = \int x$.

nonetheless underpinned this result by *contributing information* based on which the rectangle was the best combined interpretation, and in turn the occluded fourth line was decided based on that rectangle.

Somewhere in a decision-making chain there must always be[24] measurement: a certain-if-insufficient transformation from light to neuronal activation magnitudes, which will constrain assertion. For Marr (1976) the most certain assertions possible from measurement (the raw primal sketch) start this chain, by specifying input relative to possible lines and edges while unable to decide at all e.g. on rectangles. He dubs asserting only the output entirely determinable at each stage "the art of the weak hypothesis." Any *one step away* from measurement is thus maximally certain but least informative. It determines a set of entities able to somehow describe the input, but not to decide it exactly – or to organise it based on objects like a rectangle, a box or a house. Steps further removed from measurement are less certain but more useful. They can assert new sets of entities (e.g. rectangle, triangle) from entities in previous steps (e.g. line, edge) that will now be one more step removed from the certainty of measurement than their antecedents. This further step is better able to decide what is being measured through elaborating on what the previous step has already decided so far. By asserting the most basic thing decidable, then combining such assertions with rules and each other, it is made possible to assert something of greater complexity that is less certain, but of which there is now enough information available to decide at all.

This hierarchy of decisions (Marr, 1982; Palmer, 1977) is a way visual perception may overcome the limitation of any one available proximal input grossly underspecifying the distal objects that caused it according to Marr, to form "a true description of what is there" (Marr, 1982, p. 30) in the distal world. Marr (1982) thus understands visual perception as an escalation from the raw perceptual input offered by single-cell measurements, to some descriptions of objects nested in a hierarchy of increasing detail. Per the obstructed line in the rectangle from figure 2 these descriptions are not only of *visible* features, but instead present a viewer-invariant (cf. Gibson, 1979) reconstruction of the objects being perceived such that features of the description, like the fourth line in the rectangle, may not have been measured. If the measurement is reliable and the chain of connected assertions is sound, the descriptions are true: there really is an object whose features are exactly as described, that caused the original measurement via emitting light from its surface. And as a corollary, there is clearly room for error in such a process: though highly implausible, it is entirely possible a rectangle did not cause the retinal image in figure 2 and the true description of the objects is one of the cases in figure 3 instead. Error cannot be ruled out, in the form of plausible alternatives that compete (recall §2ii) with whatever description is decided on.

This tension is where the approach finally superseding the algorithms and approach from Marr (1982) connects with a much earlier line of thought. Founded equally on measurement and on assumption, it sees perception as a series of educated guesses starting not at the structure of a proximal stimulus but instead at the structure of the distal stimulus causing it. Constrained by as much information as it can get, it describes "such objects […] as would have to be there in order to produce the same impression" (von Helmholtz, 1867, §4) on the retina, by combining measured and hidden parameters to create and update their descriptions. By assuming the uncertain it can be wrong – but by only using the certain it could not describe its objects at all. Taking the story full circle, that approach is generative modelling.

iv. Success

To recapitulate: as I discussed in §2i and §2ii and then largely reiterated throughout §2iii no input for perception is specified enough to determine one individual output. Multiple inputs (or other elements) are combined to determine one or often many competing outputs. In the broader sweep of its problems

[24] I exclude visual processing entirely unprompted by retinal input, like visual imagination (e.g. Nanay, 2015).

and solutions, the computational approach I have been outlining is a match for the case I have overall advanced for visual perception – be it 'direct' or 'indirect' – as an assumption-mediated phenomenon. Marr (1974) describes assertions from sets of measurements while I have described conclusions from sets of premises. Their nomenclature aside both descriptions are entirely interchangeable as a schema, with conclusions being valid premises (discussed in 2i) and assertions being valid input for assertions. My intent with closely following Marr (1974) up until now has been to showcase how this processing dynamic of measurement and assertion, construed to make perception solvable by an algorithm of the kind e.g. Marr and Hidreth (1980) suggest for edge-detection, closely tracks an equivalent dynamic of premise and conclusion that Rock (1980) and others derive from naïvely[25] behavioural considerations. Although I will be reverting to my earlier terminology, this convergence is unlikely to be coincidental. In both cases, the demand that both seem to satisfy is the widest possible incorporation of information relevant to some measurement, whether or not the information was directly *given by* the measurement.

The specific algorithms proposed by Marr and his colleagues in the 1970s and 1980s are no longer the cutting edge. Reviewing the state of the art on object recognition, erstwhile Marr collaborators Poggio and Ullman (2013) mention the original efforts by Marr and Nishihara (1978) in passing, compared to modern work in machine learning e.g. by Serre, Oliva and Poggio (2007) or more recently Szegedy et al. (2015), He, Zhang, Rhe and Sun (2016) etc. that vastly outpaces them in speed and overall success. Past the specific case of recognition, techniques like Bayesian decision theory (Kersten, Mamassian & Yuille, 2004; Maloney & Zhang, 2010) will deliver better and faster results as a general mathematical basis for object perception based on generative models, than any algorithm suggested by Marr (1982). What nonetheless remains relevant about the approach I have discussed in the previous pages – and is the reason my focus has been on his agenda-setting essay – is in what Marr (1982) would consider the *philosophy* of his approach. The 'classical' computational theory of perception has critical elements to its structure qua theory[26] that persist in modern approaches to visual object perception even though its algorithms do not: description, and assertion from measurement revised by past and by later evidence.

First of the elements above is a reliance on description. Given my discussion of Marr throughout §2iii it should be apparent how a hierarchy from low- to higher-order inferences is utilised to 'escape' e.g. a lack of information on the distal stimulus considering only what a single measurement provided. It should also be clear that Marr (1982) took descriptions to be the output of each step – such that those inferences are always of parts, what I have called features of an object, which together can describe it. One example of parts are the edges, lines and blobs of the raw primal sketch from Marr and Nishihara (1976). An earlier example are the points, lengths and angles of orientation invoked by Palmer (1975) to construct a square. Later examples are e.g. the convex or concave curvature of surfaces, inferred by the Bayesian algorithm discussed by Kersten et al. (2004), or the nested inception modules considered by Szegedy et al. (2015) for object recognition standing for different correlating elements of an image. Setting aside differences in their implementation, and in other parts of their philosophy I will consider below, every computational method popularly applied to visual perception has revolved around using *the right parts* – whatever those parts may be, however they are found – to describe what is perceived.

The hierarchical aspect of such descriptions is simply to allow that descriptions are tiered, in levels of different distance from the most basic level of description available. In other words, that a part from a

[25] Naïve in the sense that Rock (1980) did not consider how a perceptual inference can be computed in real time and thus its efficiency, etc. Whereas Marr (1982) explicitly created his theory with implementation as a concern. The similarity of their schemata is thus a case where logical and algorithmic approaches independently coincide.

[26] 'Theory' in the 'levels of explanation' sense Marr (1982) considered them, as opposed to an algorithm. I have already invoked this usage in chapter 1 with regard to an account vs. a framework vs. a theory of a phenomenon.

given description can itself be described using parts from another. This allows escalation from a basic level to more complex levels of description till the output is determined, as Marr (1982) envisioned it. At the same time, it allows contemporary methods using 'deep' machine learning (cf. He et al., 2016) to determine images explicitly through a large number of these levels, and Bayesian methodologies to "decompose the description of a scene [into its] components" (Kersten et al., 2004) so as to align each part with one in another description generated in advance. Such approaches diverge from the classical perspective in important ways but rely at least as much as Marr (1982) on descriptions of descriptions, and they can be contrasted with reference to *how* they use such a hierarchy, rather than *if* they use one.

Description is as close as there is to a modern 'industry standard' for successful perceptual processing and of course the same is true of conceptual processing as I reviewed throughout the previous chapter. Whatever its representation, and whichever phenomenon is considered, 'meaning in the head' appears to employ descriptions just as Russell (1911; 1917) anticipates, the same way visual object perception also appears to do so. He even anticipates their convergence, taking descriptions via visual perception to form the basis for descriptions via reference as a mutually-accessible pool of predicates, in his joint account of 'knowledge by description' (Russell, 1911). This will not be my aim at present. Per Marr's art of the weak hypothesis, I will only be considering this similarity in structure qua similarity and not in the interest of convergence. It should nonetheless be clear that descriptions abound in both domains so that in both cases the aim is to align such descriptions with some "true description of what is there" (Marr, 1982) expressed in the distinct vocabulary of perceptual or conceptual representation. Whether or not the vocabulary of *true descriptions of what is there* is common between them, or is useful at all, is another matter which I will be discussing at the end of the present chapter and revisiting in the next.

The second element is assertion, and here I am being looser in my sense of both revision and evidence than is usually found in the perception literature. When I claim that Marr (1982) emphasised assertion revised based on evidence, I am making the restricted claim that he understood, in the development of the algorithms I have discussed here, that perception must eventually make uncertain claims. And that more evidence makes uncertain claims less uncertain, just how Gibson (1979) combined perspectives for what is the same sort of goal. And finally that uncertain claims may have to be revised post hoc in some cases – such that what was once an accepted conclusion, by the algorithm, will become rejected. Though a minor enough coup when spelled out explicitly, 'perception can be wrong and then be right' is an important step (as I will argue below) to grasping the secret behind its successful mechanism and it is also a founding principle of the computational approach to perception that succeeded Marr (1982)

This approach takes the idea of assertion (in the sense of claiming that there is a certain set of features – what I also called a *conclusion,* above) and starts there instead of the proximal stimulus. Whether in the form of Bayesian Decision Theory or the even more efficient technique of hierarchical 'predictive coding' (e.g. Rao & Ballard, 1999) these newer algorithms use many of the same ideas in Marr (1982) but in the opposite order. They build hierarchies of descriptions starting from the topmost features and in the case of predictive coding only record success or failure via an 'error signal', in a more technical sense than I have used 'error' here. Such approaches are varied and summarised elsewhere (e.g. Clark, 2015, 2013; Hohwy, 2013), but what unifies most of them is the focus on first modelling the expected, and then modelling how well it fits. As well as being similar in structure qua generative models, often with Bayesian algorithms, they are similar in their theoretical vision of a brain "making up the world" (Frith, 2007) and then checking to see if it got it right. Like the graphical model approach to concepts, their descriptions begin from causes and they end in predicted effects: they guess what creates a signal such that perception might be even more properly described as hypothesis in science (Gregory, 1980).

This version, or vision of perception also lets countless external influences be easily incorporated into the (per my terms) premises used to conclude features of distal stimuli. Attention, proprioception and of course language as already discussed in the last chapter (cf. Lupyan, 2012; Lupyan & Clark, 2016) are all factors known or recently discovered to feed into the outcome of visual object perception – and are easily incorporated into such ‘guessing models’, much like external factors were easy to include in similar models of conceptual cognition. Most important of all, though, for my present purposes, is that idea from von Helmholtz (1867) and Gregory (1980), under the slogan ‘perceptions as hypotheses’. In the context of generative models of perception it reflects the notion that these models are indeed given as hypotheses of the external world, tested using multifaceted (and multimodal) evidence then revised.

At the same time, it also reflects an idea that might be feasibly be adopted by (or foisted on) a broader range of theorists and theories – I would argue all those I have considered here including Wertheimer. This is that all perceptual descriptions are constructed in such a way that they are sooner or later – but always by the end of the process crafting them – presumed to be true. What makes perception tick is a positive presumption akin to null hypothesis testing, that what picture it presents is right until it is not. Perceptual illusions like the ambiguous face alone demonstrate this simple piece of bias: the picture is not unclear but rather one conclusion jostles with the other because both inferences are presumed true. It is this sense of ‘hypothesis’ true of all models of perception, and ‘truer still’ of generative ones, that relates the most to the role of perception as an enabler of contact, as I will now move to consider next.

3 Getting There

i. Baptism

I now return to Kripke (1980). In chapter 2, I considered how Kripke (1980) dismantles descriptivism from several perspectives. Then in chapter 3, I discussed how the majority of those complaints can be inherently addressed by substituting a more elaborated (and dynamic) conceptual hypothesis stated in the language of a graphical model for the simpler sort of structure originally used in descriptivism. At the same time, I also noted that one type of complaint, based on modal context, was more challenging to address in this way. And that as a result there is still room for a dedicated solution to the problem of contact, perhaps as Kripke (1980) famously imagined it. It is therefore time to see what his solution is.

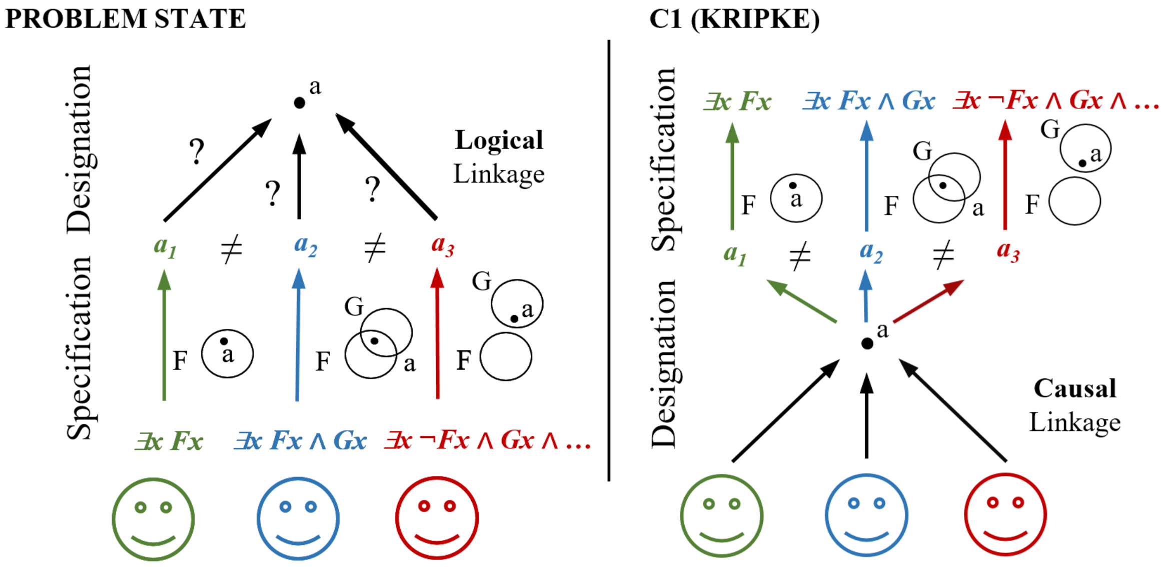

PROBLEM STATE
C1 (KRIPKE)
Specification Designation
a
?
?
?
Logical
Linkage
a_1 ≠ a_2 ≠ a_3
F a
G F a
G a F
$\exists x\, Fx$ $\exists x\, Fx \wedge Gx$ $\exists x\, \neg Fx \wedge Gx \wedge \ldots$
Specification Designation
$\exists x\, Fx$ $\exists x\, Fx \wedge Gx$ $\exists x\, \neg Fx \wedge Gx \wedge \ldots$
F a
G F a
G a F
a_1 ≠ a_2 ≠ a_3
a
Causal
Linkage

Figure 8. Logical (left) vs. causal linkage as the grounds for contact in the case of inconsistent names.

As my illustration from chapter 2 repeated in **figure 8** suggests, Kripke (1980) understood contact as preceding logical description. Instead of worrying about the ways that inconsistent descriptions could define separate logical objects linked to the labels a_1, a_2, a_3 that could not be associated with the same external object a – even where it seems we *can* associate all of them with a as in the various examples I have given, regarding Aristotle, 'Trump', 'the scientist in a wheelchair', 'the man holding a martini' – he argues the association of those labels a_1, a_2, a_3 with that object a precedes their descriptive usage. In other words, Kripke (1980) claims that designation is not a function of specification, but a separate step. Moreover, he claims that (for the types of labels I have discussed here) such designation is *rigid*.

What rigid designation amounts to is that it can be carried along modal contexts without alteration to its referent. If 'Trump' (somehow) rigidly designates the current President of the United States then it will always pick out that entity selected in our actual world, even when asking 'what if Trump were in prison', 'what if Trump were a woman', 'what if Trump was the brother of Ivanka', etc. In this, and in any other case where 'Trump' occurs in a sentence, Kripke (1980) argues the referent is constant – i.e. 'Trump' in 'what if Trump were in prison' still refers to Donald J. Trump, the President of the United States, in the actual world, who is not in prison. And a result, every problem case where a description appears misaligned with qualities of the actual referent is harmless to designation, which was set well in advance by other means, and which for Kripke (1980) *cannot be altered* by subsequent description.

The means by which designation is set according to Kripke (1980) follows one of two pathways. One pathway for designation, the one most important for this chapter, is baptism. This is where a label (by my terminology) is attached to a referent by either *defining* the referent through description or instead by *interacting* with that referent, in a way that associates it with that label. Kripke calls the latter case naming by 'ostension': some kind of pointing or gesturing or looking at or interacting with something by a cognitive agent, followed by a declaration (or just quiet acknowledgment) that from that moment hence *that* object is to be labelled 'cat' in future instances of language production and comprehension; at least instances of language production and comprehension *by that agent*, since baptisms are private.

In the case of baptism by ostension describing the information content of the entity so baptised is both valid and appropriate according to Kripke (1980): there are properties held by 'cat' when baptised that can be catalogued and attached to its label as part of a descriptive theory of (by my terms) information content. It is however not the properties that fix the referent but the ostension, that targeted interaction with *that* object (not some other, or no object) to which *that* label (not some other label) was attached: if the targeted interaction had happened with a different object, then the appropriate description would be of that object instead. If the object being baptised 'cat' was a dog the right description for 'cat' will describe a dog, specifically that dog. The right description is set at baptism, based on what is baptised. Having baptised 'cat' by ostension however, 'what if the cat was a dog' still describes the cat from the baptism, through the label 'cat', and asks what would happen if a cat of that description were a canine. Baptism therefore fixes the description: by its occurrence it provides the stamp of contact and rigidity.

The other style of baptism is baptism by explicit description. This takes the exact form of descriptions Russell (1905) would have been at home with – a logical definition of an object based on its known or anticipated properties, such as describing some astrophysical entity (cf. Frege, 1892a) by its orbit and mass and similar qualities, and then (importantly) baptising the exact entity so described as 'Neptune':

> "An even better case of determining the reference of a name by description, as opposed to ostension, is the discovery of the planet Neptune. Neptune was hypothesized as the planet which caused such and such discrepancies in the orbits of certain other planets. If Leverrier indeed gave the name 'Neptune' to the planet before it was ever seen, then he fixed the reference of 'Neptune' by means of the description just mentioned. At that time, he was unable to see the planet even through a telescope [and as a result, the content of the description is not predetermined.] Leverrier

> could well have believed that if Neptune had been knocked off its course one million years earlier, it would have caused no such perturbations and even that some other object might have caused the perturbations in its place." (Kripke, 1980, p. 79.)

As suggested by this example, where baptism by ostension attaches a description to the entity selected by other means (which will evermore be its appropriate description), baptism by description selects an entity by a description and, assuming the description is non-empty, baptises that entity so described. If 'the man holding a martini' would be so baptised then this really would refer to just some human male whose other property is holding a martini rather than water, and to no other entity. Of course, utilising the adjustments I have suggested in the previous chapter to this view of description, it is no big stretch to expand this version of baptism into the more dynamic conceptual version of information content. In his analysis Kripke (1980) nonetheless implies that this sort of baptism is an exception, rather than the rule. For most objects referred to using labels in everyday life such an intentionally scientific 'baptism by description' as suggested by the Neptune example might indeed not be how they got their meaning. Instead their referents were selected by ostension. Or they were selected using the alternative pathway besides baptism altogether that Kripke (1980), and also Putnam (1975) and Burge (1979) all argue for.

The alternative pathway for contact besides baptism is sociolinguistic transmission. One person at the start of a chain picks out the referent of a label. For Kripke (1980) this happens by baptism – which in both its varieties is only defined as a private process, carried out by one individual rather than a group sport (I return to this point below). Then another person acquires the label e.g. as part of language use. At this point the second individual did not baptise an entity using that label. Instead, so the story goes, they acquire a link to the original referent by their link to the original language user, whose label they repeat. In other words, given an initial baptism at ground level that produces some first label *l* the rest of the labels by this form *l* (recall I defined labels as tokens – so these are 'cat', 'cat', 'cat', 'cat', etc.) contact the original referent by their link through repetition, from token to token, to the one first label. I will overall return to this pathway for contact, and to its relation to coordination, in the next chapter.

ii. Doubt

Of the three means for selecting referents, or three theories of contact, given in Kripke (1980) that are all standardly summarised (not least by Kripke) under the umbrella of the 'causal theory of reference', because they are all about the primacy of connections between physical tokens like labels and entities, my focus here will be on baptism by ostension, which I will argue is ill-defined as given. I will not be arguing that the idea of rigid designation is ill-defined firstly because the persuasiveness of the modal arguments is well-accepted and passionately defended (see e.g. Soames, 2007), and secondly because it coheres with the objective nature of both human perceptual and conceptual cognition. Across these domains there is varied research on how we perceive not just an object, but *that* object, and not some e.g. alternative assembly of the same perceptual information content whose creation I have discussed.

For perception alone, work by Treisman (1988, 1992, 1998, 1999), and her colleagues (e.g. Treisman & Gelade, 1980; Treisman & Zhang, 2006), motivates a view of *feature integration* as an independent cognitive phenomenon. Based on how features connected to a particular object are processed faster in cases where they are so displayed, and processed slower when features previously bound together are assigned to different objects than before (Treisman, 1992) and similar behavioural work over decades, this is the "dominant account" (Humhreys, 2016) on how perceived features and objects relate to each other. More precisely: the account *that* features are explicitly attached to objects for human cognition. To account for this observed behaviour, Kahneman and Treisman (1984) stipulate entities whose role is to track instances of feature integration, which they dub object files. And although the exact nature of object files in terms of their links to attention (Wolfe & Bennett, 1996) and consciousness (Mitroff, Scholl & Wynn, 2005) or initial overlap against later change in location (Hollingworth & Rasmussen, 2010; van Dam & Hommel, 2010), and myriad other concerns is fluid, as with perceptual hypotheses this is a case of *how* object files are used, rather than of whether they exist at all. It is overall accepted "mid-level visual representations which keep track of persisting identity over time" (Mitroff, Scholl & Noles, 2007) do exist in some form, representing some minimal case of 'rigidity' in visual processing.

This is all further borne out by cases of neurophysiological impairment in tracking feature consistency known as Balint's syndrome (Bálint, 1909; Gillebert & Humhreys, 2015), where patients cannot build separate objects from the features they can otherwise successfully process, and cases of impairment in integrating enough features into any given object to successfully determine it (Riddoch & Humphreys, 1987; Humphreys & Riddoch, 2013). These impairments are exclusively of object (whole) rather than feature (part) processing – such that a patient in the latter case can still describe objects in exceptional detail, and even copy the objects' shape, while being unable to process how the features they describe or copy are connected. This latter case of so-called integrative agnosia (Riddoch & Humphreys, 1987) is particularly significant since the patient *can* track objects in visual space in the way that object files are intended to facilitate, but cannot assemble descriptions to visually determine objects they perceive. The type of feature integration determined here by its absence might thus be of a higher level than for object files, indicating a level of processing combining object files with more elaborated descriptions.

At the same time, this literature invokes such rigid objects specifically as grounds for revision and/or elaboration: an anchor for low- and high-level features to converge into a multi-level description of a particular entity. As originally conceived in Kahneman, Treisman and Gibbs (1992) object files track spatial feature overlap, whether features belong to the same entity in physical space. If the features in question were derived by a series of hypotheses in whatever direction, object files can contain the full chain of descriptions to 'file' objects in a more comprehensive sense. An observer may perceive a dot and then a plane (Kahneman, et al., 1992), then perhaps even Superman, seamlessly revising the same object file with all the latest information. As a result, such rigidity has a perpetually provisional sense; in the same way that perceptual hypotheses at the feature level are themselves subject to later revision. This latter part of even the most rigid of perceptual objects – that is, hypothesised perceptual objects – is one half of what motivates my doubts over the Kripkean theory of contact. The other is the extent to which ostension is feasible to start with, if this is understood to be some explicitly pre-descriptive link between objects and labels. The problem with such a view of causal linkage is that this does not work.

The contact aspect of reference is or is not a relation to a perceived, or at any rate to an intentional and individuated object. From the way that Kripke (1980) discusses these linkages, it seems to me that the 'baptism' part of his claim (again, I discuss the sociolinguistic portion in the next chapter) is meant to indeed select objects by an intentional channel: some sort of understood link between the label and its target to enable the baptism. Potentially, this could be understood as a causal connection rather than a symbolic or a logical one. And yet the selection of the connected parts itself, rather than other parts, is down to that cognitive agent doing the selecting, and their ability to select something at all to connect: even if the explanation for why this link is *reference* might be down to causality, an explanation about why or how the link between these objects, rather than another link between some other objects, came to be established is not optional. For baptism by ostension to be an explanation it needs an explanation itself. And if perception is the paradigmatic case for effecting that linkage then the way the objects are selected for baptism by ostension is itself (as I sought to illustrate through this chapter) by description. The 'ground level' view of baptism by ostension is descriptive, basing object selection on a perceptual hypothesis (or similar inferred structure) rather than any identifiable 1-to-1 linkage. That is all there is to work with – which is perhaps why Kripke (1980) concedes that "the case of a baptism by ostension can perhaps be subsumed under the description concept also (p. 96). Contact begins with descriptions.

iii. Substitution

If contact does necessarily begin with description, however transmitted later, then the question is how description can be made part of a theory of contact where phenomena like rigid designation still apply to the relationship between descriptions and referents, so descriptions are fixed to some set standard – while at the same time being the process via which the referent is identified. In response to that I will now suggest a way to turn baptism by description from Kripke (1980) into a general theory of contact most easily applicable to a generative model of perception where descriptions are explicitly of causes; – that is, descriptions that are explicitly hypotheses or theories of distal objects giving their properties. To do this I will rely on the mechanism of perception as outlined throughout the previous, for which it can be claimed (or at any rate I will not doubt here) that it typically does establish contact between the external world and our individual tokens of its objects. Going by the strength of this relationship I will then aim to set similar standards for reference: some way contact for reference may work *if* perception works; and specifically, if perception works using the sort of mechanisms I have so far described here.

First, a view of the end goal as given by C1, my earlier definition of contact for reference:

(C1) CONTACT

For some domain D and set of labels L, I take L to *contact* D, if and only if there is some criterion C based on which every member of L is assigned to a member of D or assigned to the empty set ∅.

Based on this definition, for Kripke (1980) the criterion C is causal linkage. When and only when the causal linkage exists between either a first label and an object, or subsequent labels and the first label, there is contact. Else there is no contact and the term is empty – which is also how a 'mistaken name' is determined. When there is no contact there is no reference, and contact is linkage by causal means.

What, if any, is the equivalent criterion C in perception? One way to approach this question is to use a proxy entity like an object file, and say the truth of the properties of that object file determine contact. One can do this for both perception and reference as e.g. Recanati (2012) does by recourse to 'mental files'. What is problematic about this approach, essentially descriptivism as a theory of contact (by a somewhat more empirical name), is however that object files can be wrong. As I discussed above the very point of object files, as understood by the researchers positing that construct, is tracking *change*. The consequence is not that object files are wrong, but that they are a work in progress; and the same assumption, or presumption, also underlies the other constructs of perception, features, objects and all.

Across the entire history of the study of perception, competition between notional (as for Marr, 1982), actively hypothesised (as for the generative models approach), and even active alternatives in the case of bistable images is an ever-present feature even before visual objects are constructed. Adding to this the possibility of revision after a visual object is constructed, most vividly e.g. with the Dalmatian dog but also more quietly for the birds that become planes, there is no grounds to use truth values as a test of contact. Depending on the approach to perception error is either unavoidable or a welcome part of the process of revision for any objects perceived. More than that, since (as I have argued) perception relies on assertion of what it has concluded, even where this might be wrong, pending later evidence, previously true descriptions are also insufficiently safe for use as a theory of contact. Any one could be wrong now. The key to unlocking this problem is, as I will briefly motivate below, the 'endgame'.

I have said above that perception is the contact case par excellence, as it seems to somehow get there. Except for sceptical hypotheses there is little overall doubt that perception is a successful mechanism. And what I argue most motivates this success is not the truth value of perceptual descriptions, but the control of error effected by the greater mechanism. That while description can be wrong for whatever reason, be it through conclusions (cf. the inferential terminology I have used) about features or about those assumptions being applicable, or even about whole objects being there where there is no object, what makes perception link to the world at the end of the day, is that it succeeds at the end of the day. That the processes underlying the decision-making behind visual object perception are such that they *really are*, perhaps by hypothesis mediated by feedback or perhaps using the rules-based construction Marr (1982) advocates or via whatever other means, *presumably right*. At the very end they get there and their presumption that they are true is usually justified. It is not always valid, but it is a good rule. Such a teleological (cf. Millikan, 1989; 2004) kind of view of perception does not automatically give a theory of contact for perception, or for reference. Yet it does help direct the search, away from what is claimed about any objects perceived by description, and toward an underlying trend of whether that process as a whole can be relied on to manage error: that this process is inferential (uses descriptions), but also capable of systematically fixing its mistakes wherever they occur and constrain its inferences. An achievement in which the possibility of error is not only an inevitable component but a useful tool.

In the generic inferential viewpoint, error exists in the struggle between competing conclusions, some of which are right yet only acceptable after more evidence. In the meantime there is error. In the more specific modern model of perception based on generative hypotheses, error is the essential instrument, the thing to go 'bump' against for its hypotheses to be tested and refined. Thus, a theory of perception based just on truth would be difficult to consider effective; and perhaps likewise a theory of reference. If anything links perception to the world it is its entire mechanism rather than a successful description whose success can be very short lived. The reliable success of perception given the possibility of error is the more valuable element – that it is good at correcting its mistakes or accounting for new changes.

What is important to underline is that, in the absence of the right mechanism to constrain it, falsehood easily becomes destructive. It is only as a function of a mechanism capable of revising false cases into true cases that error becomes an integrated element – and truth values (arguably) a secondary concern. The converse is once again intuitive to demonstrate by a pathological case, the joint pathological case of perceptual hallucinations and delusions. Perceptual hallucinations and delusions (Frith, 2005) form a success story for cognitive neuropsychiatry in explaining belief-related symptoms for schizophrenia, brain trauma and for other pathologies, and even hallucinations in healthy individuals, (e.g. Teunisse, Cruysberg, Hoefnagels, Verbeek, Zitman (1996). Such hallucinations manifest as "false perceptions" (Fletcher & Frith, 2009). They are a "sensory percept in the absence of a genuine stimulus" (Corlett, Frith & Fletcher, 2009), or a perceptual equivalent for 'mistaken names': perceptual description of an object with no role of input at all. The other side of the coin is perceptual delusions: irrational beliefs held in defiance of available sensory input equivalent to descriptions that are being arbitrarily revised.

These cases of altered perception are informative because their now-canonical modern explanation by Frith (e.g. Frith, 2005), Fletcher (e.g. Fletcher & Frith, 2009) and Corlett (e.g. Corlett et al., 2009) has connected them to the balance of (in my terms) asserting descriptions, and revising descriptions based on input. Although correctly seen as success cases for the forward-modelling paradigm, what seems to create the problem could be stated more generally, as an inability to combine old and new information (modelled in the form of prior expectation and feedback). For hallucination prior assumptions on what *may* be the objects perceived override the input on what objects there actually are. While for delusions accurate perceptual conclusions may become arbitrarily revised into inaccurate ones. This all usefully illustrates the consequence of error when revision is impaired in either direction, both in being unable to control the information of descriptions, and being unable to adjust these descriptions to fit the data. What could be most taken home from these examples is that perception might be able to follow a path from assertions that are often wrong to assertions that are often right, but does so as a function of what allows that progression to be positive. Where falsehood is benign that is predicated on the mechanism. Yet where falsehood really is benign, because the process either has modelled or is going to model an accurate facsimile of the distal stimulus – particularly in generative models where this is the only sort of content present – then there is room for a more pragmatic Criterion C: representation in that model.

The best way to understand this suggestion is to consider the ‘endpoint’ for some idealised instance of perception, where all objects that are present and could have been modelled by a cognitive agent have been modelled accurately, and where no objects that are not present have been modelled by that agent. For the sake of simplicity and comparison to reference I will assume this model is generative in nature such that the objects are inferred causes of the proximal stimulus, and they really are the causes of the proximal stimulus. In that case, a criterion of contact where contact is effected just by true description would conclude that contact is made. Every one of these objects is contacted and their descriptions by which the model has gotten to them could therefore be fixed in some perceptual equivalent of rigidity. What I am now suggesting is that, if the average output of perception is close enough to its idealised output (and I have assumed that it is), then the same criterion can be applied. For this imperfect case there is no guaranteed connection from truth to contact which is why I do not consider truth to be an appropriate Criterion C. Even in a great model of a target object there will very likely – as a function of the inverse problem – be objects some of whose features are inferred incorrectly. Nonetheless, the majority of what is given by the model will be true. As a result, being represented by that model will be a suitable approximation for being part of the true model. Contrasted to the purely causal criterion of contact given by Kripke (1980) this might instead be considered the *statistical criterion* of contact.

In other words, what I am suggesting is that perhaps the standard of contact should not be individual entities, but the robustness of the entire mechanism by which they are delivered. Where that model is indeed, by the process of its construction and maintenance, a good fit for the target phenomenon, just the way perception is a good fit for its objects, then just being part of that model is itself a criterion of contact. Or counterfactually: if perception is indeed successful on a routine basis, only the entities that exist and only the correct properties of such entities are likely to have been represented by perception. If an entity does not exist (in the visual field) or does not have those properties, then the model would not hypothesise its existence – or it would be revised to exclude it. Though what I suggest amounts to placing a bet on that model, what I am also suggesting is that this just might be a sure enough gamble.

4 Stealing Contact

In this chapter, I have considered the process by which perception establishes contact with the objects we presume to perceive. Through the course of this discussion I have underlined the inferential aspect of this process and its reliance on description to forge even the simplest sort of link. I have considered the fundamental joint role of description and revision in mustering and maintaining these connections.

Finally, I have considered the Kripkean theory of contact by causal linkage, argued that it reduces into the problem of contact for perception and then argued that the problem of contact for perception could be solved as a function of the robustness of the underlying mechanism. Where the mechanism crafting the model of objects perceived is robust enough, I have argued any asserted output may be substituted for the ideal output, and so *being an output* of perception is sufficient for a statistical contact criterion.

Chapter 5: Worldcraft

1 Assembly Required

> “The world is everything that is the case.”
>
> (Ludwig Wittgenstein, ‘Tractatus Logico-Philosophicus’, 1922, §1.)

Through the past three chapters I have sought to motivate, first the questions defining the pieces of the philosophical problems of contact, content, coordination for reference, then plausible accounts for two of those pieces – content and contact – combining empirical mechanisms with philosophical concerns.

In particular, I have argued reference can be understood as this triple problem from how the questions asked by the three pre-eminent philosophers analysing reference through the late 19th and 20th century (regarding the generality and commensurability of reference, the information content of empty names, and the inaccurate predictions of a descriptive and truth functional theory of designation) may each be analysed individually; under the respective guise of the problems of coordination, content and contact. I then considered how the particular problems raised in light of those questions can be addressed, first for content, in the form of a psychological theory of description casting concepts as graphical models; then for contact, in the form of a statistical criterion assuming perception is a firm way to establish the desired link, and capitalising on its robustness against falsehood. As with the questions, my arguments were all in isolation from each other. And I postponed issues related to coordination for either context.

In this chapter I finally bring these separate strands together into a combined framework, aimed at the combined triple problem of contact, content, and coordination for reference. To begin with, I consider how the criteria of contact and content established from chapters 2 and 3 could carry over to this more complete picture, and I motivate thinking of reference as analytically underwritten by *notional worlds.* As my next task, I then consider some remaining general problems or concerns related to coordination to show how they could be adequately addressed first in a private context and then in a public context. I lastly consider two objections along opposite strands related to the view of reference I advocate here, and how these may be convincingly addressed via the merits of an effective theory of coordination: in particular, a theory facilitating transmission of both contact and information content through dialogue.

Overall, the theory of reference I present here is (as I have noted previously) a framework, rather than a fully specified formal or computational system. Much of what I discuss is based on ideas elaborated by other disciplines, like the graphical models approach whose details I left vague; or they call for an extended programme of research to substantiate, like the connections between concepts and dialogue. Nonetheless, exploring the initial potential for dialogue as an effective interface for coordination does take up the next chapter, in the form of a brief overview of results from two experiments that examine effects of conceptual categorisation and feature inference in the course of cooperative dialogue games.

2 Worlds

i. Creation

First of the three answers to bring together, in order of numbering and in order of importance, is C1:

(C1) CONTACT

> For some domain D and set of labels L, I take L to *contact* D, if and only if there is some criterion C based on which every member of L is assigned to a member of D or assigned to the empty set ∅.

Something is needed that can make this assignment for reference, and in the last chapter I argued that perceptual representation, as a feasible version of the criterion of baptism by ostension, can fulfil that role. In other words, that being represented at all by perception (in its necessarily descriptive way that I discussed in the previous chapter) can suffice as a theory of linkage, where that linkage is a function of some cognitive agent doing the baptising (who must 'point it in the right direction'), and a referent. Shorn of the empirical discussion of perception entirely, the reasoning behind this move is as follows: firstly, some form of robust link is required between a label and some object that serves as its referent. Secondly, based on the modal objections raised by Kripke (1980) it emerges that logical description is not a robust enough provider for that link. Specifically, truth values for descriptions are insufficient to guarantee or (as important!) to reject the necessary link, and referring to an object by description does not seem to transfer across modal contexts; or at least I accept here the Kripkean view that it does not. As a result, my next step was to assess the suggestion in Kripke (1980) that this sufficiently rigid link can be established by ostension or by scientifically-minded description, as part of an original baptism; and then transmitted by virtue of the chain of usage from that original baptism to any later token for a label. An original baptised 'cat' (or 'Trump') thereby underwrites every subsequent instance this way.

Yet because the scientifically-minded description is quite unusual, and because the ostension case will reduce to a claim that cognitive agents can select a target for their linguistic behaviour via some direct method, which is incompatible with any empirical theory for perception, another answer was required. To preserve the idea that perception is sufficient to select a target for baptism, or otherwise underwrite contact for reference as also suggested by other authors (e.g. Fodor & Pylyshyn, 2015), my next move was to consider in what form perception might be able to offer a substitution criterion. I rejected plain descriptivism based on perceptual rather than logical descriptions as a criterion for contact, for similar reasons to Kripke (1980) having rejected descriptivism in the first place. Namely, that the truth values of perceptual descriptions were insufficient guarantees of whether the description finds an exact target when taken for individual cases. Instead, I suggested that perception was robust enough that any ideal outcome (whose truth values will be exactly accurate) and most actual outcomes (some of whose truth values might not be exactly accurate, but most would be) are close enough to exchange for each other.

Given this 'statistical' level of equivalence in the two cases, and the important proviso that perception tends toward truth and away from falsehood over time as a function of its mechanism, I then argued it would be productive to consider *representation in such a robust model* as a sufficient criterion for the individuating intentional link between a cognitive agent and an object, I took Kripke (1980) to mainly call for with his 'baptism by ostension'. In other words I argued that perception can select a target just by representing one, albeit by description, because its representation-via-description is highly reliable: when there are objects to represent and only then, it represents them, and it does so accurately overall. If adopting this as a theory of contact for reference, the criterion linking a label to a referent would be its perceptual representation, since that representation (in the form of a description) suffices to find it; it is not the fact that this is perception that was important but rather that this can individuate a referent.

Presently I consider another option for satisfying the criterion for contact based on the same reasoning I gave for perception, that if an individuating mechanism is robust enough, then representation by that mechanism (by description and/or a belief in descriptive form) is sufficient to select a referent – albeit a candidate where the critical provisos of robustness, and tendency toward truth are more conditional. The alternative I have in mind (perhaps predictably) is conceptual, rather than perceptual descriptions. Prima facie the individuating descriptions in the case of perception and the individuating descriptions in the case of conceptual cognition (i.e. those used to identify a novel instance of some category) will not be very different kinds of things. In particular, if the perceptual description is taken to come from a generative model where the descriptions are hypothesised features while the conceptual description is taken to come from a generative model where the descriptions are hypothesised features, then what separates the cases is *the type of feature* (visual vs. conceptual) and, more importantly, the robustness of the underlying mechanism. Setting aside robustness (which is the main worry), there is good cause to think the two sorts of features are not fundamentally different. They are certainly similar enough to interact, so that e.g. salient perceptual features affect explicit description by conceptual properties in a ‘feature listing’ task (Wu & Barsalou, 2009; see also Barsalou, 1999; Lupyan & Clark, 2016) – so that even intuitively the idea that both of them feed into some greater pool of properties that are tracked by cognitive agents with regard to objects in their environment is not far-fetched given their connections.

What instead separates perceptual and conceptual cognition, with regard to their potential as guarantor of accurate individuation (of the type that a baptism by ostension would require) is robustness most of all: that perception is demonstrably accurate in most cases, given the *exceptional* character of illusions or hallucinations, whereas conceptual cognition is not as obviously accurate, or as biased toward truth. Conceptual models can be generally wrong; they can be misguided with no real effort to correct them. At the very least, some argument or evidence is thus necessary to suggest that conceptual descriptions can ‘stand in’ for the objects themselves they describe. Though if they can stand in this way, the same argument I gave for perceptual descriptive representation being a suitably strong proxy for the objects themselves ought to then also extend to conceptual representation. I will now offer an argument that a case can be made for extending this substitution argument to the conceptual case – but as a function of context. In other words, that conceptual description could possess this robustness *for certain contexts*; and that as a result it is possible to extend the strategy I have pursued for perception, in these contexts.

Two equally important steps drive the argument, adapted from work by Dennett (1987), and Millikan (1989; 2004). The first of the steps is that we could *capture a snapshot of every belief* some organism (in my terms, some cognitive agent) has formed over their environment. The result will very much be a subjective construct, in the sense that there is no attempt to evaluate these beliefs. Whatever beliefs are present, including those representing biases or other inaccuracies, they are taken to be true: that is, for the sake of the argument, it is assumed this cognitive agent is correct about all they believe. Then the second step for this argument is the implicit assumption, that *the beliefs held by a cognitive agent are typically conditioned by their environment* (in the long-term sense; i.e. their evolutionary habitat).

A cave-dweller has extensive and detailed beliefs about light and shadow. A frog-eater has extensive beliefs about frogs, which the cave-dweller does not. That is: both the range of targets and the quality of such beliefs will reflect what the cognitive agent has spent its time doing and/or learning about and (the key step for my usage) it is also more likely than not that their beliefs are evolutionarily adaptive. If a cognitive agent living in an environment, of whom we took the snapshot, has a belief or does not have a belief, this is typically part of their adaptation. The past history of that agent and perhaps of its species will have had some impact on shaping its beliefs. In fact: if this snapshot is all true, the agent is perfectly adapted (belief-wise) to an environment of which these beliefs really are all true – except where false beliefs are adaptive, which I will not consider here. Moreover, this environment of which their beliefs are all true could, in turn, be *described using the agent's beliefs as a guide*, by taking the full set of beliefs as a definition of just what is the case within that environment; their *notional world*.

This construct, the notional world, is a world based on the beliefs of a cognitive agent assumed to be adapted to an environment. It thereby reflects a specific viewpoint into that environment, conditioned by the agent and their interests: so e.g. a frog-eater notional world would have very complicated frogs. A cave-dweller notional world will have barely any sort of frogs – maybe just fish that can sometimes jump. To be clear, this will be because the world is being *constructed* according to the blueprint of the appropriate belief set, and a frog-eater has very detailed beliefs about frogs, whereas the cave-dweller does not have detailed beliefs about frogs and as a result frogs would be very simple for their notional world – the world constructed exclusively and exhaustively with cave-dweller beliefs for its blueprint. For these notional worlds, *what is believed to be true* and *what is the case* are one and the same thing. Dennett (1987) presents this device to make the point that we may analyse beliefs *in abstractio* of the entities and properties they are beliefs of (whether true or false), while preserving the level of analysis whereby e.g. beliefs affect behaviour relative to objects in an environment. Whether or not the objects are accurate these notional worlds nonetheless capture what the cognitive agent would think of and do with their habitat: if they believe frogs can fly, it captures the way they behave relative to these flying frogs; whether or not they are 'real' flying frogs. Setting aside the overtly theoretical or philosophical use of these notional worlds, however, there is a somewhat different potential application to entertain.

If we grant for the sake of the argument that a cognitive agent is perfectly adapted to an environment, in the way that teleological accounts of reference most notably assume is possible given enough time and where the environment is not changing (see e.g. Millikan, 2004) – not a notional world where the adaptation is by hypothesis, but *an actual environment* in which the cognitive agent has been thriving, then the notional world of that cognitive agent gives something slightly different. In being a snapshot of their beliefs about the environment for which the agent is perfectly adapted, while their actual, real environment *is* the environment for which the agent is perfectly adapted, their notional world offers a picture view of the actual environment the agent lives in. More than this, it gives a view, from within that environment, of what the agent has beliefs about: an ontology of the things with which it has had important transactions. If there are trees they climb in their environment, the notional world has trees, and these trees have leaves and branches. If there are trees in their environment but the agent is some forest-floor forager then perhaps the trees have no branches in the notional world – because there has never been a use for them. This teleological notional world thus paints a true but partial picture of the real one based on what the agent would have needed to have beliefs about and only that set of entities.

Moreover, if the beliefs of the cognitive agent take the specific form of e.g. a graphical model making hypotheses about entities and properties, which attempt to capture *the properties of the actual entities these beliefs concern* (as opposed to sense-data – cf. my discussion across chapters 3 and 4), then that agent's notional world will also be a partial reconstruction of the (actual) objects in their environment. When an agent's beliefs are formed as hypotheses, like perception, their corresponding notional world will be a putative description of what there is. It would include all those entities that the agent has had interactions with and will exclude those entities (in the actual world) that the agent has not considered vital enough to understand. In some computational sense, this will *be* a world simulating those entities an agent has a history with and (again) omitting those entities the agent is either unaware of or has not had to form beliefs about. This notional world thus selects between the two kinds of entities, by either including them in its 'simulation' or not. What is simulated in the notional world is something that has had to have true beliefs formed about it for adaptive reasons; and what is not simulated in the notional world is everything else – things an agent is historically unaware of. And as a result, where the beliefs might be labels, this assignment to either set of beliefs precisely matches the role for criterion C in the definition of contact: a criterion sorting between those things labelled, which therefore have some link to the cognitive agent important enough to represent and label them; and those things for which there is no contact, because there is no label (or corresponding belief) for them at all, in the notional world.

To unpack this slightly: what I claim here is not just that notional worlds built of beliefs in the form of generative models – the exact form of 'information content' I discussed in chapters 3 and 4, for which 'belief' here is effectively a synonym) sort between objects with labels that refer and those with labels that do not refer. I am suggesting that labels for things that do not refer will not be part of the model at all: that *they will not have labels without referents simulated in the first place* – since, if they had been important enough to simulate, then they would not be empty. The criterion thus maps to my definition of contact, by ensuring that exactly the entities with labels are also the ones for which the labels ought to refer – in turn, because every labelled entity that is represented in the model is an entity with which the owner of the model has had a transaction of the kind that allows them to label it. Each label would exist only due to an appropriate individuating causal linkage (i.e. a baptism by ostension) in its history when an agent picked that object out, since (again), had they never singled that object out, they would have formed no beliefs about it. And had they never formed beliefs about an object, that object would not be a part of their notional world; and it could not have been labelled by the agent in the first place. What holds this argument together is the promise that this generative model producing an appropriate notional world (hereon: World) can be trusted to be accurate enough that, what it does not model will not be relevant; and in the other direction, all it does model (and label) is accurate for its environment. *For models satisfying these assumptions*, representation of labels in the model is itself the Criterion C.

As a result, assuming there is only a single cognitive agent in that environment (and therefore no links to referents through peer labels, as I discuss immediately below), and only so long as the environment in question is stable to the point where beliefs about its entities can reach a maximally accurate status, it is possible to give a theory of contact for referents relative to an environment where *the availability of information content* for a label, including minimally the label itself, alone implies some prior valid instance of baptism by ostension conducted in that environment. And so that any information content attached to a label implies information content attached to a referent. The form of this argument may suggest some sleight of hand has taken place. It has not: my claim is ultimately an elaborated version of a claim explicitly made at several points by Kripke (1980; 2013) that mistaken names, i.e. the hard problem case of contact, will be eliminated given long enough time. He does not explain how or why except to suggest users of mistaken names would retract them, if or when they realise they are empty. And for that idealised environment I have been discussing, all users eventually must. More precisely: they will either realise their beliefs over the labelled object are not productive to keep (after however much time has passed in that constant environment), or a label will be unused for so long they forget it, eliminating it from their model. Explicitly or implicitly adaptation will have corrected their World.

Clearly enough, unlike perception this solution is not generic. Like the special pressures hypothesised to form stars, any Worlds the membership of whom is a criterion sufficient to address the problem of contact need special circumstances to be created. Yet there may be an easier way to make them work.

ii. Curation

The above leaves a twofold objective. Firstly, to consider how to make this contact-solving analytical structure, a World whose membership is Criterion C, apply more broadly than the limit case from the evolutionary example. Secondly, to assess how the definition of content given by C2 can be met by it:

(C2) CONTENT

For some language $\mathscr{L}$ and set of labels L, I take the *content* of L to be the specification of all L in $\mathscr{L}$.

Happily, meeting C2 for these Worlds, constructed as they are from hypotheses in a generative model information structure, is simple. As suggested in chapter 3 one can take a graphical model of concepts and their properties and use the model to generate input-free hypotheses for each concept. Though the exact form of these hypotheses will vary based on the more exact 'species' of concept – e.g. prototype vs. exemplar; cf. Danks (2014) on how these would each look – it is still possible to give a hypothesis for each concept type; which for labelled concepts (i.e. the ones covered by C2) will include the label.

Since the presence of these concepts and their labels in the model implies that they *are* linked to some object whose category they capture (due to C1; and recall here that I have considered proper names to be just single-object categories), each label will be specified in the language of graphical models via a description that can stand for that object itself. That is: each description will also be the corresponding definition of an object from the World (the model) by which the information content is being attached. And by extension, if C1 is met, each description will be the definition of an object in the actual world. This last consequence of appealing to conceptual, rather than perceptual, descriptions to fix reference, namely that contact based on concept definitions automatically defines content through the very same concept definitions, and vice-versa, is important to keep in mind for the bigger picture of this solution.

At the same time, it is helpful to keep in mind that appeal to Worlds, in the strict evolutionary context where environments are stable enough to make conceptual information tend toward accuracy, is not at all the only way to approach this sort of framework. Perceptual description would also fix the referent. And it could very well provide that link by which conceptual hypotheses are attached to objects under *any* circumstances. In fact, the need for a special context mainly depends on whether one further thing is true or not. If every item with a label has a perceptual description attached then perception performs the exact same role as conceptual description in the above: for every labelled entity given a perceptual description there will be (by the 'statistical' criterion) some target entity to which the label is attached. What creates a problem for this alternative is the possibility of purely conceptual description where an item might have a label but may never have been perceived, and so its contact would come into doubt. Though I will return to my reasons for considering this particular issue, it may well be that the best-fit framework for capturing contact, content and coordination ought to employ perception for criterion C.

The other alternative pathway to consider for this solution is via the salience of context for Worlds. In the last section I have already emphasised the importance of a stable context, wherein the information in a World will become close enough to the real properties of real objects, that substituting the former for the latter is feasible. Yet the particular information structure of graphical models – as I previously discussed regarding 'the man holding a martini' – has the advantage of keeping track of not only one context at a time (for example: one environment) but several contexts toggled by an appropriate node. In the fuller picture of cognition where context effects are rife this suggests that such a representation could supply 'a toggle for every occasion': wherever concepts are applied there will be an adjustment made to the ways their features relate to each other and to novel objects, e.g. if Casasanto and Lupyan (2015) are correct about concept models. (Though if they are correct no context can count as 'stable'.)

Moreover, as notional worlds (and so Worlds) are defined explicitly *for an environment*, such context toggles imply there can even be multiple notional worlds derivable from the same background source of information. And this seems correct. There can be some stable context, e.g. a household, where the objects are constant enough that they can acquire maximally accurate information content (if they are relevant enough to know well); and yet the same can be said at a broader level of 'environment' such as a city, whose landmarks are stable, but whose household-level objects will be constantly changing. For the city what will be stable and thus a candidate for the sort of real-to-notional substitution-based framework that I am advocating here, would be the buildings and the parks. For one household it may be the toothbrushes and individual inhabitants. Different entities may be stable enough to replace for each of the levels of description considered, if their stability is the necessary condition for the switch. More generally, it is possible that smaller subsets of environments can be stable enough to satisfy C1 by the substitution method I gave in chapter 4, even where the overall environment is a poor fit for it.

And thus, even where contact could be assured by some other means, this curation of notional world models per context – be it different times, locations or levels of description – is a vital feature. It will identify items and labels (among the sum contacted) stable enough to track or not *in this context*; and thereby which items should be included in a notional world used to analyse reference for that context. Crucially, this context sensitivity is inherently delivered by the above models of conceptual cognition.

It is rather important to keep in mind the phrasing, 'notional world *used to analyse* reference'. What I am at heart suggesting with this talk of 'Worlds' of reference is not that there is some additional thing on top of how we represent information content for reference where these substitutions and the like all internally happen and are represented as mental events. To be clear, what I am suggesting can be done just like Dennet (1987) suggested in his conception of notional worlds: an existing cognitive structure, which I have argued is already effective in capturing the information content of reference, will just be analytically thought of as underpinning the question of contact, by this series of possible substitutions, under the right circumstances. I am not suggesting there is any substitution or similar manipulation in any sort of implemented way. It is *the analytical possibility of the substitutions* I consider here for C1. My further argument regarding context only elaborates material from chapter 3 from a different angle.

Setting that caveat aside, stable-context Worlds implying contact (even just as objects of analysis) are a useful tool for constraining analysis of symbol grounding and expectations of when such an analysis will be viable through the notion of substitution. This might be (as suggested above) per environment. But it may also be per concept, as different concepts can be more or less stable both by their usage in inference (e.g. Malt, 1990; Gelman & Markman, 1987) or by their ambiguity (e.g. Hoffman, Lambon Ralph and Rogers, 2013). Perhaps notional world analyses might apply for stable subsets of concepts, as I have already suggested they could apply for subsets of stable parts of a greater environment. And again, this is only an argument for considering *each pre-existing model of content* as a notional world, and thereby a solution for contact as well. There is no additional cognitive mechanism that I appeal to.

iii. Connection

Lastly, there are some preliminary points to make on the third problem for reference – coordination:

(C3) COORDINATION

For a set of labels L, I take reference to be *coordinated* between cognitive agents, if and only if by each label in L the agents i) designate the same entity, and ii) specify it with the same information.

Although coordination is most meaningful for the case of agents interacting with each other, using a label to mean the same thing, there is a reduced account that can be given without agents interacting based only on what I have covered so far. This concerns the case where two agents occupy instances of the exact same stable context at different times. If these agents were to attach information to what labels they use (and by extension their referents), expressing hypotheses exclusively about what is in that context, based on information exclusively derived from that context, this is a 'solipsistic' model for how reference coordination can be achieved by 'learning the same lessons from the same entities' – without actually communicating or in any way comparing information content or contact conditions.

For this ideally stable environment, the same one where I have argued the C1 and C2 criteria are one and the same, there is therefore no need to give an additional account of coordination, except for CS:

(CS) COMMENSURABLE SPECIFICATION

I call two specifications S_A by cognitive agent A and S_B by cognitive agent B, of referents r and q, *commensurable*, if and only if i) S_A may be substituted for S_B with no change in the information A attaches to r based on its label, and ii) r = q.

Where the information two cognitive agents each attach to a referent could be swapped between them, these agents were said to satisfy condition CS, the background assumption for reference coordination; and one not as easily by the minimal example I have just given. For any language used to describe the objects in question to be identical across the cognitive agents in question, an additional connection is required. I consider CS more in the next chapter, since – for reasons I consider there – taking on this question is not easy to do either by observation, or experiment,. It must be answered by assumption. So, although important to keep in mind throughout, and concerning if it were false, I will assume CS.

With CS secure until later on, a theoretically helpful aspect of this description- and hypothesis-driven take on reference is its capacity to elaborate on the sort of linkage suggested by Kripke (1980) and by Putnam (1975) and Burge (1979), which in its traditional reading entirely transcends the ground level (i.e. psychological processing) analysis of how 'cat' got its name – via appeal to historical connection and sociolinguistic transmission. As I outlined in chapter 4, this family of theories can be summarised by the device of of causal linkage through iterations of use. My own label 'cat' will inherit its referent from someone else's, or a whole community's, prior label 'cat'; of which my use is the latest instance. And this analysis is effective, in that there is indeed good reason to think that examples such as 'gold', 'tiger' (Putnam, 1975), 'Moses' (Kripke, 1980), or 'arthritis' (Burge, 1979) are often used in a starkly underspecified manner. And as a result the information content encoded by 'tiger' is properly defined via some baptism equivalent (notwithstanding my definitional objections for baptism) yet it might not be accessible to a cognitive agent using the term. In contrast to whoever baptised the tiger, subsequent users can *refer to it correctly* using bad information – so this is a case of contact without specification.

The transmission of reference through a wider community of language speakers, i.e. a 'sociolinguistic community', is essential to the appeal of externalism. Without it, each user of reference is an island in their own right, able to causally interact with the world and putatively create chains of reference from their labels to the objects they interact with. Yet much like the isolated organisms evolving in a stable environment I considered as all my examples so far, they are trapped in their immediate environment; and their ability to refer does not extend beyond their own limited experience. It is only by leaning on other language speakers that these isolated individuals may expand their ability to refer – reaching out across time and space to tigers, Moses, and everything in between. For the externalists, this might not always (of even often) mean that any useful information is accessible to subsequent users of the labels (such as how much a tiger weighs or what Moses looks like) but the transmission is a story of *contact*.

Yet such externalist theories for what permits reference to travel from user to user without being fully specified are both prima facie unsatisfying, in thought experiments like Swampman (Davidson, 1987), and also to an extent mysterious, in that they begin at the point where a new user has already acquired e.g. the label 'cat' without properly accounting for the mechanism of its acquisition from another user. As with baptism this abstraction appears harmless enough: *language happens,* then a label passes to a new member of their sociolinguistic community. Philosophers need not really delve into the fine print. To be abundantly clear, these are philosophers displaying enormous care, attention and intelligence in their work, whom I am not accusing of negligence. I am rather considering the division of labour they had envisioned, between the analysis of reference, and (what I argue is not) its trivial implementation.

What is left out by this level of abstraction is the real possibility in line with all the empirical work by Lupyan and others that I have discussed, that every new user independently reassigns a novel referent for the labels they inherit based on their own understanding – even as part of the very communication necessary to transmit it, as I will argue with certain recent results further below. With mounting work for effects of labels as input, even on their own, not just in solidifying but in prompting and directing categorisation, a perceived case where communication only transmits – or only importantly transmits – contact without content is tenuous. This not because content can matter to individual members of a sociolinguistic community after the fact (which no externalist denies; they deny its links to reference) but because the chain of transmission is itself 'tainted' with conceptual processing linked to language learning and production. As long as a label is transmitted, some information is transmitted along with it, and some of that information enables the learning required for transmission in the first place. Much like the argument I gave in chapter 4 against the false simplicity of identifying the target of baptisms, defining *the transmission appealed to* in discussing sociolinguistic community label transfer demands an analysis of content, if only to define what was transferred, and whether it was the expected label at all, or an unrelated homonym, deviating vastly from its intended usage and repeated by pure accident. In other words, 'lining up' the labels being transmitted begets further explanation, most plausibly via conceptual/language processing ('meaning in the head') vs. exclusively causal or metaphysical means.

Even if appealing to instances of transmission for describing contact does not itself involve content it is thus infeasible to eject content from how labels are successfully transmitted, unless the 'labels' are understood as physical vibrations created by a speaker and replicated by the next; by which point any initial understanding of 'sociolinguistic community transmission' would have become greatly warped. An externalist view of label transmission is thus valuable as an abstraction but omits crucial detail for a more exact analysis of the progression of contact and/or content across iterated label tokens. That is not to say the externalists are wrong about sociolinguistic transmission; only that there is no easy way to omit the information content of reference from how labels are transmitted. And if transmission is in turn what underwrites contact for labels, then 'contact by transmission' is tainted by content past what seems reasonable to abstract from, assuming a more modern scientific understanding of language use.

Yet the upshot of these sociolinguistic connections between labels being harder to describe without an account of content (and at that an account of content informed by scientific understanding of language use) is that *having* an account of content of the sort already available provides an opportunity to finish the picture painted by externalism about how reference is transmitted across language users. (Albeit in an adulterated way.) Specifically, interactions between language users can become more well-defined by considering how they operate on different levels of representation (e.g. lexical, conceptual), whose properties and referents can be altered even by implicit language use – as I discuss in the next chapter. A scientifically informed view of language users as messy, clashing communication viewpoints could thus supplement, rather than supplant the original externalist approach to sociolinguistic transmission. And this account can in turn be developed into a fuller (future) theory of conceptual coordination, one defining the processes by which language users can come to (in my terms) share their World, just like those in the 'ideal' minimal example who never meet; so long as they share their underlying language.

In particular, the hypothesis that representations across all levels of structure, from phonetics to (more importantly) conceptual content, can become *aligned* during joint language use suggests an immediate possible avenue of investigation, as I will explore in the next chapter – both for understanding what is being transmitted when labels are explicitly taught and learned, and for exploring how language users incorporate passive language input to alter their conceptual understanding. As language can 'program the mind' (Lupyan & Bergen, 2016), by affecting categorisation more directly than thought of before, there is even room for asking if a conversation could be a source of input comparable to perception in its impact on our concepts and knowledge; and therefore our successful coordination of these, as well.

For now, the take home-message is that although I do not give an account of coordination in this or in the next chapter (though I do sketch the first steps toward one in the latter), a framework for reference couched in the terms I have outlined above is well-poised to offer one, e.g. by elaborating externalism into a more hybrid theory of causal transmission guided by information content in its implementation; just as I previously argued for elaborating baptism into a hybrid theory guided by information content.

3 Objections

I have now discussed at some length the idea that reference may be productively analysed as a whole, using a framework where the putative output of generative models of conceptual information content designates referents in the equivalent notional world (World), to the point of substituting the original referents as the object of analysis; be it in ideally stable environments or the more approximate cases.

Although the above discussion, more so than the three major chapters preceding it, was couched on a positive premise – of 'what if we solved reference this way' – two informative avenues of objections are still worth considering here. These represent different extremes and can help bring the framework I have discussed into clearer focus by what they each demand and, in turn, omit to achieve their aims.

The first avenue of objection is reductive in its perspective, represented most clearly in a very similar speculative view of reference by Fodor and Pylyshyn (2015) where instead of the various notions and problems of information content and coordination, all that is used to explain referential phenomena is perceptual causal linkage. In other words, why not purely appeal to perception? To be clear, this view is not externalist. All of the analysis I offer for perception in chapter 4 (bar the substitution argument) is typically repeated in this minimalist explanation of reference – explicitly so by Fodor and Pylyshyn. Where this approach deviates from the one I suggest is in doubling down on perception as determiner of reference – a role which I have already argued even in this chapter, it is very capable of playing on its own. Taking perceptual representation for contact, perceptual information content for specification (e.g. the description of geometrical shapes in space that I considered in chapter 4) and even perceptual coordination (in the sense of e.g. identifying the same objects from the same perspectives) as the three criteria for the problems I attribute to reference, one can get the advertised 'minds without meanings'; and therefore an analysis free of the more particular complications of concepts, and their organisation, although even a 'purist' analysis of perception requires effects from e.g. prior beliefs to be articulated.

What such an account brings to the table most of all is the robustness of perception as a guarantor for contact, which the World framework attempts to isolate and replicate for concepts in certain contexts. In the purely perceptual case the robustness afforded is the real deal, with no further need for concern. What this account omits, however, is the reach of any account also coached in concepts and language.

And this is purely down to the expressive and inferential power of using the fuller model. Conceptual effects like feature inference allow concepts to have much richer content than just those given by one (or several) instances of perceptual linkage. Although there is certainly overlap in the information that perception and conceptual processing attempt to recreate, the loss of robustness by taking concepts as the cue for designating referents (rather than perceptual descriptions) is counterbalanced by the many more properties – including many influencing perception – this richer, less certain model can capture.

Moreover, the potential for a theory of *coordination* trading on conceptual (vs. perceptual) content to explain and incorporate the transmission of reference within language communities will again expand the number of referents a conceptual analysis can encompass, compared to a perceptual analysis built literally on what the eye can see, and no more. And last but not least, much of the 'complexity' of the conceptual approach I have advocated – where interactions between concepts are explicitly modelled – is not, as Fodor and Pylyshyn (and like-minded authors) worry, potentially explosive in scope. With graphical models (or a similar structure) explicitly in mind as the form in which this information may be represented, the various rules for relevance between nodes I discuss in chapter 3 exist precisely to curate a smaller set of what *most matters* for each concept in the model, and so each labelled referent.

The second avenue of objection is from the other extreme, and much harder to answer. This approach asks whether the linkage between contact and content is productive – that is, how does the framework gain anything compared to an externalist analysis of contact and some separate analysis of conceptual cognition for the benefit of reference? And it asks its question in the shadow of an important theorem, formally demonstrating any set of descriptions (*qua* predicates) can be assigned arbitrarily to referents despite any possible constraints over whether these descriptions are true or not (Putnam, 1981). In fact this objection is aimed almost precisely at the sort of picture I have tried to present. Putnam starts, as I do, with the notional world analytical construct from Dennett (1987), and the same evolutionary angle of attack I have pursued here, on how the notional worlds can ostensibly converge with the actual one. He then uses this (effectively) *logical double dissociation* between descriptions calibrated using some notion of accuracy or truth, and possible configurations of referents for these descriptions, showing in the starkest possible logical terms that unless there is an inherent connection between them other than truth, the internal consistency of the former (in my terms the consistency of information content) will never necessarily predict the corresponding real-world objects (in my terms, the referents designated).

In a nutshell, this objection not only denies rule CS for coordinating referential language between any humans but a form of rule CS for coordinating reference between any systematic human language and its putative referents. If so, there would be no reason to attempt to unify these elements. Moreover, the premise of my Worlds argument is too uncertain to form a viable framework – mirroring as it does the exact premise of Putnam's imperious deconstruction. As a result, the answers I consider are – perhaps inevitably – partial and more speculative than logical in their nature. I will nonetheless offer two here.

One answer is to suggest that enough of this combined framework concerns how *human beliefs* about the objects surrounding them are internally curated and potentially transmitted that, even if ultimately misdirected, understanding the delusion is a useful effort. That there is some intuitive appeal to every part being involved in an analysis of reference, rather than provocatively emphasising that some parts of reference perceived as useful by its users, such as its information content, are ultimately perfidious. That is: even if the project is doomed as a logical analysis, it may be useful as an anthropological one.

Another answer, the one I opt for here, is to concede that some inherent link is necessary, between e.g. conceptual content and certain kinds of referent, to 'bootstrap' the framework. Attempting to sketch a path to such an answer closely relates to rule CS, and coordination – that I now move to consider next.

Chapter 6: Concepts and Coordination

1 Philosophical Preliminaries

The previous chapter gave the titular framework for reference, but also concluded on a troubling note: despite the tidy way the elements of the framework fit together, under the appropriate constraints, the logical double dissociation by Putnam (1981) hangs heavy over its potential to fully unlock reference.

The present chapter sketches a future path from two important parts carried from the previous chapter. The first is the potential for inherent links between conceptual descriptions, including labels, and their putative referents. The second is the beginning of a theory of coordination, elaborating the externalist conceit of sociolinguistic transmission for reference, and uniting this with the Fregean considerations prompting my discussion of coordination in the first place: trading notes, to agree about what there is.

These two approaches may seem related to different problems, but they connect intuitively as follows: if there are any inherent links between certain elements of conceptual descriptions and some referents, and I will sketch an argument that there are, and if the Fregean perspective of coordination as a quasi-scientific endeavour of comparing concepts and attempting to reconcile them is accurate, and I again will suggest that it is, a key objective for such a process would be securing these important concepts. That is: the process of coordinating reference will not just focus on having e.g. *the same* descriptions but on having *the right* descriptions, offering the best chance of passing any 'logical dissociation test'.

For this to be possible there needs to be some evidence of inherent links between any human concept and some specific information about the actual world that the concept will express. And I will further narrow this to inherent links between human concept *labels* and specific information – assuming as I have above that labels are concepts. Having narrowed the question thus the answer is yes, a strand of work on very early language acquisition has found consistent links between primitive lexemes (labels) and the referent these are taken to stand for. (Monaghan, Shillcock, Christiansen & Kirby, 2014; Imai & Kita, 2014). Such connections are very basic in nature – yet are found in both linguistic analysis of monosyllabic labels, where similar labels tended to refer to similar objects, e.g. 'dog' is closer to 'cat' than to 'gear' for both the label and referent in question (Monaghan et al., 2014), and also in naming and in world-learning behaviour by children and indigenous cultures (Imai & Kita, 2014); extending well beyond ideophones like 'bark', and into conventionalized words, like the names of slower or of faster birds being themselves longer or shorter, and comparable implicit predicate-level descriptions.

There is therefore good reason to think that, hard as they are to track beyond these deliberately basic samples, inherent features of labels themselves can map to respective inherent conceptual features of objects themselves, leading at least one group of others to specifically celebrate solving the word-to-referent association uncertainty problem (Imai & Kita, 2014). Whether or not this solves the problem outright, the existence of such connections does suggest some parts of conceptual models can survive even the logical double dissociation challenge by Putnam (1981). Moreover, if our selection of labels stems at least partly from the underlying structure of their referent, this also motivates assuming, as I have in the last chapter, that rule might be CS sufficiently accurate to assume it could be true overall. That is: humans seem to inherently share at least *some* parts of their descriptive language in the form of labels encoding, and facilitating the learning of, similar implicit information via their pronunciation and form. As a result, although CS cannot be proved or observed for some entire descriptive language, it it is nonetheless plausible to claim the building blocks of human conceptual descriptions are similar.

2 Experiment 1 – Introduction

Our aim with the two sets of results summarised in this chapter was to explore a key question raised over the previous analysis, elaborating the role of language as a medium for conceptual coordination by asking if a conceptual coordination mechanism comparable to the natural scientists trading notes envisioned by Frege (in his 19th century view of natural science) is active in online dialogue; 'online' here in the technical sense, of happening *as language is being processed* just by automatic means as opposed to happening when explicitly considering what one said, and the information they provided.

That is to say: we investigated the extent to which simple task-oriented dialogue settings, of the kind explored by Krauss and Weinheimer (1961), then revisited by Clark and Wilkes-Gibbs (1986) at the dawn of the modern study of human dialogue, serve as an interface for conceptual coordination. The typical expectation for similar language games is that specific lexical or syntactic processing may be elicited, in a way that changes the players' observed performance or behaviour, relative to a baseline. Accordingly, our expectation here was that specific *conceptual* processing – distinct from the simple manipulation of lexical labels, e.g. to shorten them – will be elicited as part of these language games, in a way that alters the players' observed behaviour. In both our experiments the focus is specifically on demonstrating that conceptual processing in its most typical forms of *categorisation* (for this first experiment), and *feature inference* (for the second set of results), could be elicited online as part of a collaborative language game, in a way that affects its outcome against the predictions of other factors.

The significance of doing this relates to the overall explanatory device of *alignment*, widely accepted in psycholinguistics as the underlying drive for human interlocutors, whether automatic (Pickering & Garrod, 2006; 2013) or deliberate (Clark & Wilkes-Gibbs, 1986), to match each other's language use. When interlocutors are e.g. syntactically aligned, they will use similar syntax. When they are lexically aligned they will use similar labels. When they are conceptually aligned they will use similar concepts (e.g. Brennan & Clark, 1996). A variety of non-linguistic factors, including beliefs about interlocutors (e.g. Branigan, Pickering, Pearson, McClean & Brown, 2011) influence *lexical alignment* – the use of identical words (in our case specifically *labels*) by participants negotiating a task together. And so we asked whether *conceptual* alignment (more accurately *mis*alignment) will influence lexical alignment.

For the collaborative account (e.g. Brennan & Clark, 1996) lexical label choice is contextually driven by agreement – a 'conceptual pact' – between interlocutors on a shared label reflecting some common conceptualisation of the referent having entered their 'common ground'. For the competing interactive alignment account (e.g. Pickering & Garrod, 2006), label choice is driven by mechanical repetition of recently processed language structure – perhaps using forward modelling (Pickering & Garrod, 2013). The collaborative account of label choice predicts that players may initially debate a label and concept but once agreed will then retain the convention even across different rounds of the same dialogue task. The interactive alignment account predicts that when interlocutors use the same labels they will come to privately conceptualise their referents in the same way too: but if so, this conceptualisation process is vulnerable to interlocutor-internal interference from each interlocutor's other concepts, and perhaps creates such interference for these concepts as well, connecting label choice in dialogue to conceptual coordination. In this first study groups of three participants played a picture-matching game (Clark & Wilkes-Gibbs, 1986) in successively interacting pairs, with the Matcher in a round becoming Director for another player in the next round; an arrangement adapted from Garrod & Doherty (1994). We then manipulated whether a ‘Learner’ player – initially acting as a Matcher and then as Director – had seen and pre-conceptualised the experimental pictures, abstract forms known as ‘tangrams’, by being asked to privately conceive appropriate descriptive labels for them. The Learner thus played the game either having previously conceptualised and conceived possible labels for the experimental tangrams, or not. Moreover, the Learner player was inserted into the game specifically such that their first round would cast them as Matcher, receiving descriptions from a Director player, which they might not agree with.

Our overall expectation was that when a Learner was privately asked to conceptually categorise their pictures by describing them, their private labels would not always match the ones previously used in the game, and as a result later rounds in which the Learner was introduced to the game will stand out from the others, being overall slower and resulting in more changes to the labelling conventions used.

We further expected the conceptual salience of labels should increase players’ tendency to use them – and that Learners who had preconceived labels will *increase* the mean label length measured in later rounds, against the well-recorded tendency for mean label length to *reduce* as the interaction proceeds (the result of players becoming more efficient and their locutions more refined after successive plays).

We finally elicited private descriptions from all players, expecting that the labels they have aligned to use in-game, will in turn predict the private conceptual descriptions they provide following that game; and where the labels all players used in a game were aligned, their post-game labels will align as well.

3 Method

i. Participants

54 native English speakers were recruited from the University of Edinburgh community in 18 groups of three, and paid for their participation[27] ($\bar{x}_{age} = 19.2$; 70.5% female). 3 more groups (9 participants) were recruited, though excluded from analysis: two due to excessive participant error with following instructions and one due to experimenter error. It was important that groups establish a reliable set of common labels, for private labels to contrast with, based on very few iterations of our language game. To make this easier, the members of each recruited group were all previously acquainted, so that their mutual knowledge (Clark & Marshall, 1981) and prior conversational history (Wilkes-Gibbs & Clark, 1992) could both act in favour of the lexical alignment we intended to elicit. Participants still entered the study without any prior conceptual grounding – common or private – for the experimental stimuli.

ii. Materials

Software: Participant dialogue was recorded in its entirety using the ‘Dialogue Experimental Toolkit’ (DiET; Mills & Healey, 2013). DiET was built to record live text-based (vs. spoken) natural dialogue, and facilitate automatic experimenter intervention (cf. Mills, 2014; Mills & Healey, 2006). Our use of it was, however, restricted to recording participants' utterances: an appropriately configured text-based instant messaging interface can functionally approximate spoken dialogue for the purpose of language games like the present (Mills, 2014), while removing the room for errors in voice-to-text transcription. For our particular configuration of DiET, we limited the number of dialogue turns on visual display at any one time to the most recent two: the last two entries typed by any player and followed by ENTER. This meant that at any given time participants could access just the last two things said – a limit meant

[27] Where eligible, participants were compensated with course credit as an alternative.

to simulate working memory constraints in spoken natural dialogue, by requiring participants to recall earlier parts of their conversation, rather than easily copy and paste relevant parts of their chat history. Memory-based constraints likely underpin the collaborative conventionalisation behaviours we aimed to generate with this experiment (Reitter & Lebiere, 2011), so it was important that we preserve them.

In all other respects, this was a familiar instant messaging environment. Each new entry was privately available to the person typing it until made public by pressing ENTER; and an indication was present of whether another player is typing, reflecting the norm for major commercial proprietary chat clients. This particular feature made our chat client strictly dissimilar to spoken dialogue, where interruptions typically occur mid-utterance (cf. Clark & Schaefer, 1989), and cannot 'unsay' what was already said. Nonetheless, its resemblance to commercial chat products allowed frictionless participant interactions with our client given their experience with the former. We therefore chose to maintain that familiarity over strictly simulating spoken dialogue at the cost of a more unusual instant messaging environment.

Stimuli: Two sets of 35mm x 50mm printed image cards were used for the communication game part of the experiment. The first ('target') set consisted of 4 tangram figures from Clark and Wilkes-Gibbs (1986) and 2 tangram figures created for this experiment. Our novel tangram figures were assessed by an experienced tangram puzzle solver for their geometrical complexity and cohesion with the rest and found consistent in both complexity and visual style. The full set of six tangrams is shown in **figure 1**.

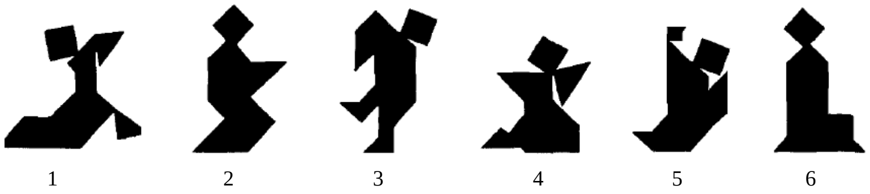


Figure 1. Tangram figures: target set. Leftmost four from Clark & Wilkes-Gibbs (1986).

We selected these particular tangram figures from Clark and Wilkes-Gibbs (1986), and supplemented them with comparable ones of our own making, for their moderate visual complexity and the multiple ways they can be visually interpreted as a result. Our aim was for figures with multiple interpretations that are nonetheless still intuitively *interpretable* – as e.g. "the sinking boat" or "karate kid" (item #5). The second set consisted of the same six figures plus a seventh tangram (not pictured) from Clark and Wilkes-Gibbs (1986). This seventh 'foil' tangram forced participants to label each of the six tangrams they had in common – those in our target set – in order to, at least initially, disambiguate from the foil.

For the private conceptualisation task (pre- and post-interaction, as described below) one A4 sheet of either the six tangrams in the target set, in the same 35mm x 50mm scale, or an entirely unrelated set of six animal tangrams – omitted – was headlined by the question "*What is depicted in each image? Please give a brief description.*" Together with our verbal clarification of this instruction (as below) our aim with the private stimuli was to prompt participants to *conceptualise* these items as they felt was right, as opposed to just *describing* them to an imaginary audience (cf. Fussell & Kraus, 1989).

iii. Procedure

Setup: Upon arrival participants were briefed on the general rules of the matching game from Krauss and Weinheimer (1964) and Clark and Wilkes-Gibbs (1986) we adopt in this experiment: two players each have a set of image cards in front of them concealed from the other player's view, and each play different roles in the game. The Director player has their cards arranged in a random order at the start of the game and cannot move them. The Matcher player must then arrange their own set of (identical) cards in the same order the Director started the game with. With neither player having access to both sets of cards, Director and Matcher must collaborate to arrange the cards exclusively using language, typically by the Director describing each card and the Matcher identifying their identical counterpart.

In our specific version of the game the Director and Matcher were each in a different room with a set of cards and a computer terminal. The Director had the 'target' set of tangram figures (see Materials) and the Matcher had the second set with the seventh additional tangram. Director and Matcher had to therefore explicitly identify, and then arrange those six tangrams they had in common, into a random two-by-three pattern given to the Director at the start of the game. There were no further instructions or restrictions on how players could describe the items to each other, or on the correct strategy to use. The only stipulation was that they use our chat client running on the nearby terminal to communicate. All participant inputs to the chat client were recorded on a remote server connected to both terminals.

Having explained this two-player game to them, each group of three participants were then told about how they would *take turns* playing it across three separate rounds, as part of our 'daisy chain' schema. The way this was explained to them was that a Matcher in a given round becomes Director in the next round, with the previously-idle player coming in as Matcher, until all players have been both Director and Matcher. This is an accurate description, though we will return to this schema immediately below. Finally, the three members of each group were randomly assigned to one of three (permanent) roles in the game: Player A, Player B and Player C, whom we will dub the 'Learner'. No advance information was given to participants about the private conceptualisation task, or our interest in the Learner player.

Schema: Our overall three-round chain (where a round is one or more games with a constant Director and Matcher pairing) was designed to first *elicit* collaborative labels for the target stimuli, in an initial round, then to *transmit* and refine these labels in later rounds, within what Caldwell and Smith (2012) call an "experimental microsociety". Our aim was for labels created collaboratively in Round 1 by the initial Director and Matcher to filter down into later rounds where one or the other initial player is not present, and where we can manipulate the cognitive state of the player being substituted in their stead.

To achieve this, we adapted a method used by Garrod and Doherty (1994) to simulate transmission in the context of a sociolinguistic community by having participants interact in *interchanging pairs* until each participant interacted with every other participant exactly once, evenly sharing their conventions. Garrod and Doherty (1994) found that overall coordination of participants increased for interchanging (as opposed to constant/isolated) pairs with every new round. This cumulative strengthening of jointly established conventions, robust enough to inspire later work e.g. by Garrod, Fay, Lee, Oberlander and MacLeod (2007; see also Garrod, Fay, Rogers, Walker & Swoboda, 2010) on whether collaboratively strengthened conventions can explain the emergence of symbolic representation, was exactly what we needed to ensure a stable culture of collaboratively-created labels could be established as our baseline. Another way to understand the same idea is by the 'replacement method' (Mesoudi, 2007; Mesoudi & Whiten, 2008) for simulating cultural evolution across generations. Generation 1 creates a convention, then some members die and others are born, leading into a Generation 2 with mixed membership: part of Generation 2 are individuals carried over from Generation 1 while another part are novel members. As a result of this overlap cultural conventions can accumulate from one generation to the next as the old interact with the new in real-world and experimental (micro-)societies (Caldwell & Millen, 2008).

Round 1: Our chosen adaptation of Garrod & Doherty (1994) took place across three rounds. Round 1 comprised a single iteration of our matching game (as above) with *Player A as Director* and *Player B as Matcher*. Though one iteration was not expected to be sufficient to establish *stable* labels Player

A and Player B nonetheless *collaboratively* established initial labels for the tangrams – which usually also entailed an initial conceptualisation of the tangrams as e.g. a jug (#4), boat (#5) or gargoyle (#3).

At the same time, the player in the Learner role performed the *private* conceptualisation task, writing down six brief responses for "what is depicted by each image" (see Materials) for either the target set or a foil set of unrelated tangrams. In addition to the written instruction, we verbally clarified that the question was "like an inkblot test: just write down the first thing that comes to mind in 3 to 5 words" and presented the task as unrelated to the main experiment. As a result, no Learner player mentioned the private conceptualisation task to any of their subsequent interlocutors until the final debrief stage.

Round 2: Following the establishment of collaborative and – in the Preconceived condition – private Learner labels for the target set of tangrams in Round 1, along with any associated conceptualisations, Round 2 was where collaborative labels initially established in Round 1 were transmitted and refined. For this we extended the length of Round 2 to span three iterations of the matching game played with the same players in the same roles throughout – recreating the conditions for collaborative refinement of referential expressions *repeating* in a dialogue setting described in Clark and Wilkes-Gibbs (1986). Since lexical repetition in a dialogue setting is the primary driver behind lexical refinement according to Clark and Wilkes-Gibbs (1986; cf. also Pickering & Garrod, 2004), there needed to be enough of it.

The pairing used in Round 2 was *Player B as Director* (from Matcher in Round 1) and the *Learner as Matcher*, with Player A stepping out of the game. As a result of the switch, the labels used in Round 1 had the chance to be transmitted by Player B to the Learner when Player B as new Director used them to describe the tangrams. This is also where our experimental manipulation first became relevant, as a function of *how much the Learner knew* when they entered the game. Learners in the Naïve condition entered Round 2 with *no prior labels* or conceptualisations for the tangrams, whereas Learners in the Preconceived condition had *labelled the tangrams before* as part of the private conceptualisation task.

In sum, Round 2 consisted of Player B directing the Learner across three consecutive iterations of our matching game. And the Learner either first learned labels for the tangrams from Player B or they had already created labels for themselves in advance of the game. In the meantime, Player A was kept in a separate room, unable to follow the game (cf. Wilkes-Gibbs & Clark, 1992), until Round 2 had ended.

Round 3: Following the establishment of collaborative labels for Round 1 and their transmission and (predicted) refinement for Round 2, Round 3 closed the loop with the *Learner as Director* and *Player A as Matcher*, while Player B stepped out of the game. This complete 'daisy chain' loop – depicted in **figure 2** – from Player A directing Player B to the Learner eventually directing Player A was an exact match for the stipulation from Garrod and Doherty (1994) that each player take up each role just once.

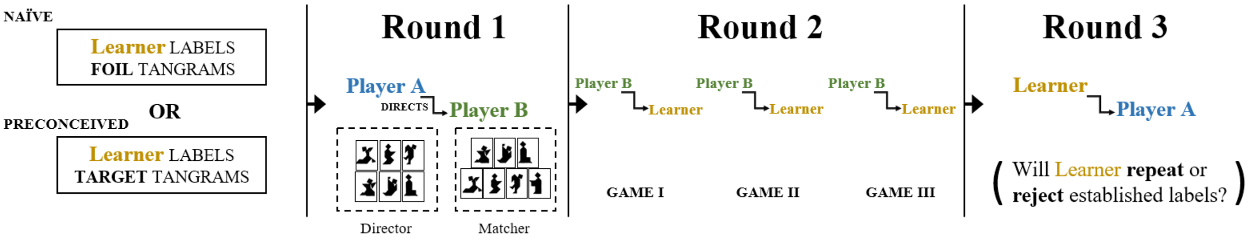


Figure 2. Visual summary of the three-round 'daisy chain' matching game, our manipulation and measurement.

Post-interaction: Following the communication game portion of the experiment, all three participants performed the private conceptualisation task for the target tangrams. For Learners in the Preconceived condition it was the second time they performed this task and they were allowed to alter their answers.

4 Results/Discussion

ii. Measures

Our basic measure was *conceptual consistency* between labels in two instances of the matching game: a binary measure of whether the initial label produced by the Director in a game (such as Round 3, or the third game of Round 2) matches another initial label produced by the Director in a different game (such as Round 1, or the first game of Round 2) in the manner it conceptually categorises the tangram. If two labels are conceptually consistent, they conceptually categorise the same thing in the same way.

The exact form of the measure was more precise and summarised immediately below. It will be useful to emphasise in advance, however, the contrast between measuring conceptual consistency (as we did) versus *lexical* consistency – a more traditional measure with which ours partially overlaps. It will also be useful to clarify our measure of (pairwise) 'consistency' contrasted to broader 'global consistency'.

With regard to conceptual vs. lexical consistency, two labels would be lexically consistent under the strictest possible definition, if and only if they use the same words: such that e.g. "seal with a ball on its nose" is not a lexical match for "seal with a ball on its *head*". This strict definition, from Brennan and Clark (1996), and related work in the same literature (e.g. Brennan, Schumann & Bartres, 2013), serves to restrict room for interpretation of what counts as similar. However, it also underdetermines the scope of the difference between using "seal" and e.g. "dancer" as labels for a given (#1) tangram. Since our main objective here is to study the divergence between one convention for describing (and by extension conceptualising) a tangram figure, and an *alternative* convention, our criterion for what counts as consistent was less sensitive to minor lexical deviations like that between 'nose' and 'head'.

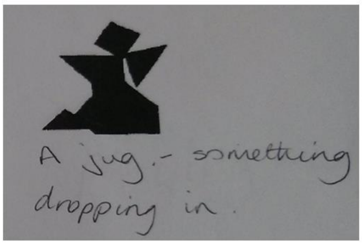
We then derived two binary metavariables tracking conceptual consistency across multiple label pairs: cumulative consistency, and post-interaction alignment. *Cumulative consistency* was scored 1 if and only if every recorded label pair for a tangram up to and including a given round was conceptually consistent. For example, if a label from Round 2 Game III was scored 1 for cumulative consistency this meant all labels from Round 1 and Round 2 Game I (all past games on record – since we skip Round 2 Game II) were both consistent with Round 2 Game III, and they were also consistent with each other. If a label from Round 3 was scored 1 for cumulative consistency, it meant labels from every other round in the game were all consistent with Round 3 and they were also all consistent with each other: the tangram was described using just one convention. If a Round 3 label was scored 0, the convention was broken.

In this latter case cumulative consistency is also a measure of alignment across the game. Cumulative consistency was scored 1 at Round 3 if and only if all label pairs in the game were consistent; making a separate variable (e.g. 'game alignment') unnecessary. *Post-interaction alignment* on the other hand was computed separately, and coded whether post-interaction labels for a tangram were conceptually consistent. It was scored 1 just when a tangram was privately described by Player A, Player B and the Learner all using the same convention. For both cumulative consistency and post-interaction alignment a score of 0 did not indicate at which point the relevant 'chain' had broken, or if the chain was broken more than once.

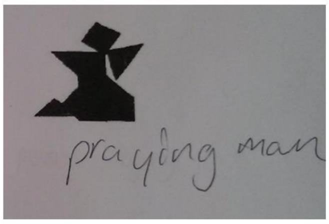
Tracking the presence or absence of established conventions (via the series of conceptually consistent labels represented by cumulative consistency) helped us to precisely test our hypothesis that Learners in the Preconceived condition would be more likely to break with established conventions, measuring whether a global convention exists as opposed to whether any specific round locally matches another. Likewise, tracking the presence or absence of conceptual alignment at the game- and post-interaction level allowed us to explore whether within-game alignment (measured by cumulative consistency for Round 3) predicts post-interaction alignment (measured by its dedicated variable) for each 'microsociety' as a whole.

Beyond conceptual consistency, we also recorded *label length* measured in words, including function words (e.g. "the") but not including hedges (e.g. "looks like") and *game or round duration* measured in minutes, beginning at the first recorded input by either player and ending at the last recorded input. Lastly, we used *animacy* (VanArsdall et al, 2015) as our purely conceptual measure of label salience.

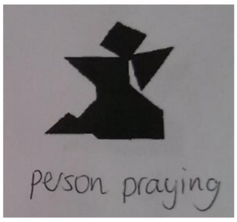
ii. Qualitative



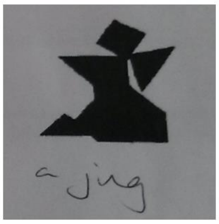
Learner Pre-game



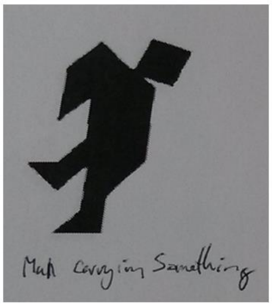
Player A Post-Game



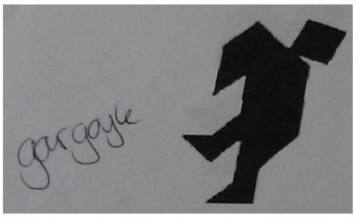
Player B Post-Game



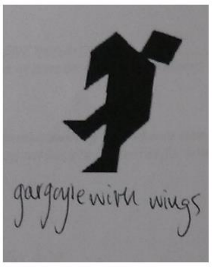
Learner Post-Game



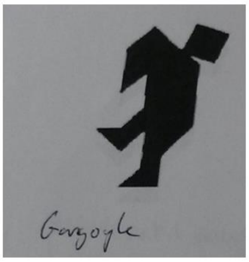
Learner Pre-game



Player A Post-Game



Player B Post-Game



Learner Post-Game

Post-interaction Preconceived Learner labels *are* often revised (vs. pre-interaction) to match a stable convention. Yet cases where they are *not* argue against an uncomplicated alignment interpretation of label choice, considered alongside cases where the convention is broken post-interaction by Player B.

iii. Quantitative

Duration: Mean game duration (in minutes) over Round 1, Round 2 Game I, Round 2 Game III and Round III are summarised in **Figure 5** across the Naïve and Preconceived conditions. As anticipated, duration was significantly reduced from the first to the last game in Round 2 ($\bar{x}_{t\ 2\mathrm{I}} = 11.16$ min, SD = 5.06, vs. $\bar{x}_{t\ 2\mathrm{III}} = 2.09$ min, SD = 2; t(17) = 7.33, p < .0001) compared using a paired-sample t-test.

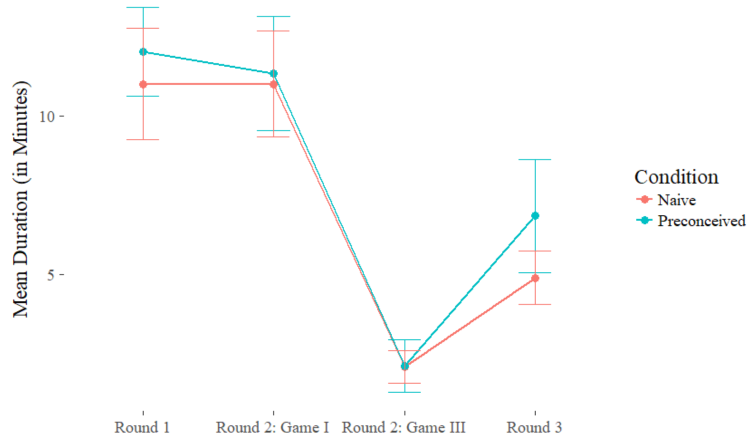


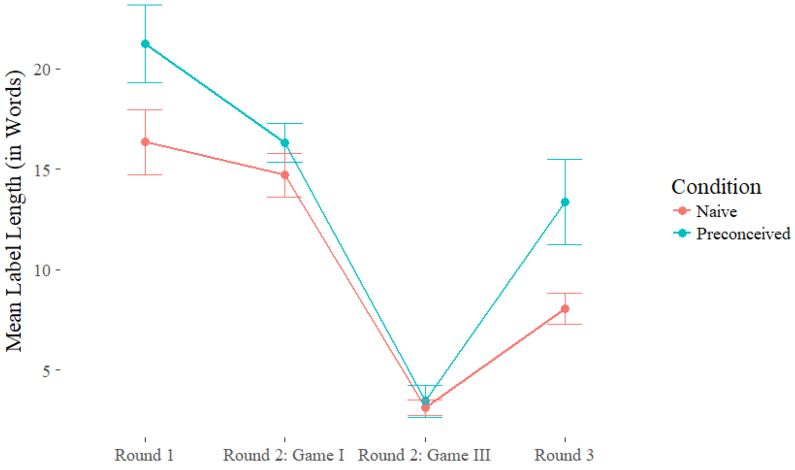
Figure 5. Mean matching game duration in minutes across conditions.

At the same time, mean duration in either Round 2 Game III or Round 3 was not significantly different for the Naïve against the Preconceived condition. Mean duration in Round 2 Game III was virtually identical between conditions ($\bar{x}_{t\ 2III\ Naive}$ = 2.06 min, SD = 1.57, compared to $\bar{x}_{t\ 2III\ Prec}$ = 2.12 min, SD = 2.47; t(16) = -0.04, n.s.). Mean duration in Round 3 was overall longer for Preconceived groups ($\bar{x}_{t\ 3\ Naive}$ = 4.88 min, SD = 2.52 compared to $\bar{x}_{t\ 3\ Prec}$ = 6.84 min, SD = 5.33; t(16) = -0.99, n.s.) but so was the standard deviation. Our hypothesis on mean game duration across conditions was thus not supported.

Refinement: Mean label length (in words) across Round 1, Round 2 Game I, Round 2 Game III and Round III are summarised in **Figure 6** across the Naïve and Preconceived conditions. Like duration, label length fell significantly from the first to the last game in Round 2 as we had anticipated ($\bar{x}_{l\ 2\mathrm{I}}$ = 15.5 words, SD = 7.6, vs. $\bar{x}_{l\ 2\mathrm{III}}$ = 3.28 words, SD = 4.48; t(106) = 14.3, p < .0001) when compared by paired-sample t-test. This is consistent with our overall prediction for label refinement in Round 2.

ditions.

Mean label length in Round 3 was also significantly higher in the Preconceived condition ($\bar{x}_{l\ 3\ \mathrm{Naive}}$ = 8.05 words, SD = 5.77, vs. $\bar{x}_{l\ 3\ Prec}$ = 13.35 words, SD = 15.4; t(105) = -2.37, p < .01) compared by independent t-test. Again, this is consistent with our prediction about Round 3 Preconceived Learners.



What was less expected was a difference between conditions for Round 1 before the effects of our manipulation. Round 1 labels for the Naïve condition were significantly shorter than for the Preconceived condition ($\bar{x}_{l\ 1\ \mathrm{Naive}}$ = 16.33 words, SD = 11.72, vs. $\bar{x}_{l\ 1\ Prec}$ = 21.24 words, SD = 13.66; t(101) = -1.96, p < .05) compared using a two-tailed independent t-test. We have no hypothesis linked to this effect – yet its presence may have impacted the Round 3 data. We revisit the matter of interpreting Round 3 label length in light of Round 1 in the discussion.

Consistency (General): To assess label consistency we used mixed-effects logistic regression models constructed in a uniform way. Our models were fitted via the lme4 package (Bates, Maechler, Bolker & Walker, 2015) in the R software environment (R Core Team, 2013), using the 'logit' link function. We opted against using a maximal random effect structure (cf. Barr, Levy, Scheepers & Tilly, 2013):

this would require group and individual item as random predictors for both intercept and slope in our case but we used just *group* as the random effect included in all models and only as random intercept.

This decision was motivated by two practical concerns. First and foremost, any random slope effects would be unlikely to converge in a model where each of 18 groups only supplies 6 total observations, and likewise, each of 6 items only supplies 18 observations (given we only considered single rounds). Secondly, for individual items, our procedure did not involve recording the correspondences between labels and individual item in a systematic fashion, since the arrangement part of the task was physical and the typing data was recorded off-site. As a result, it was not possible to conclusively match every label to a specific tangram for every group. We therefore excluded an item variable from our analysis.

For our fixed effect structure we used forward selection, starting from the random-effects-only model and adding predictors in the order they are mentioned in the text. Every new predictor was retained if and only if model fit (assessed via likelihood ratio test) improved *and* either the predictor itself or any of its interactions with existing predictors (assessed in that order) was significant. Any nonsignificant predictors or interactions were removed from the model, before moving on to the next listed predictor.

One label for each of Round 2 Game I, Round 2 Game III and Round 3 was not available for analysis, as a result of three different players ignoring the foil when arranging the tangrams in later rounds and therefore only labelling (in these cases) the first five. No pre- or post-interaction labels were missing.

Our main hypothesis was that the experimental manipulation would raise the likelihood an established convention would be abandoned by two different stages of the game: Round 2 Game III, and Round 3. **Figure 7** presents cumulative consistency across conditions for Round 2, 3 and post-interaction labels.

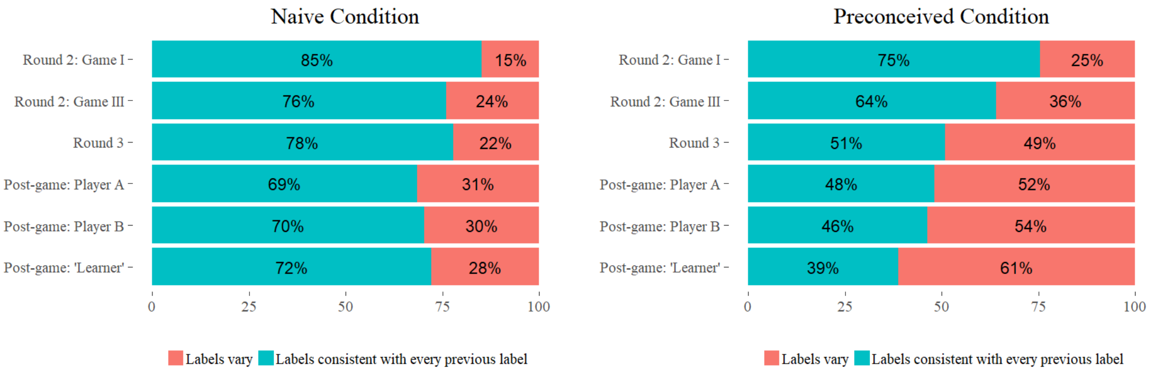


Figure 7. Cumulative consistency across conditions for Round 2 (Game I & III), Round 3, and post-interaction labels.

The overall pattern of cumulative consistency suggests descriptive conventions introduced in Round 1 were quite likely to be maintained across the entire game, especially in the Naïve condition. In Round 2 Game III (the first predicted ‘break point’) 76% of labels remain cumulatively consistent for Naïve groups, compared to 64% in the Preconceived condition, down from an initial value of 85% and 75%. In Round 3 the proportion of cumulatively consistent labels remains constant for the Naïve condition; the apparent increase is due to a missing value – whereas for the Preconceived condition the share of cumulatively consistent labels drops precipitously to 51%, suggesting our manipulation had an effect.

We include cumulative consistency for post-interaction labels to illustrate the extent that conventions persisted even after the dialogue was over, transferring from a collaborative context, for the matching game, into the private conceptualisation task. Taking Round 3 as a baseline there is only limited drop in cumulative consistency: 69%/70%/72% of post-interaction labels for Naïve and 48%/46%/39% for Preconceived players follow an earlier established convention, down from 78% and 51% for Round 3.

Consistency (Round 2): When predicting the likelihood of cumulatively consistent labels in the last game of Round 2 there was no significant effect of condition, despite the proportion of cumulatively consistent labels being reduced in the Preconceived condition. There was a significant main effect of consistency with the post-interaction label produced later by Player A (β = 1.184, SE = 0.56, p < .05). Although there was no main effect of consistency with the post-interaction labels produced by Player B or by the Learner (in isolation), there was a significant interaction effect, for consistency with both at once (β = 2.284, SE = 1.11, p < .05). When the label from the last game of Round 2 was consistent with post-interaction labels produced later by both the Learner and Player B, it was much more likely this label would follow the convention established going into Round 2 rather than break away from it. There was lastly no significant effect of label animacy on cumulative label consistency for this round.

Our hypotheses for Learner interference and conceptual salience for Round 2 Game III were thus not supported by the model. It was not significantly likelier an earlier convention would be broken for the Preconceived condition, nor did that likelihood vary with label animacy. However, consistency with a post-interaction label later produced by Player A reduced the likelihood that an established convention would be being broken in Round 2 Game III, where Player A was absent. And though post-interaction labels by either player present had no impact alone, consistency with a post-interaction label produced later on by both Player B and the Learner (such that Player B and Learner post-interaction labels were also consistent with each other) vastly reduced the likelihood an established convention was broken in Round 2 Game III. These effects have opposite bearings on our 'conceptual entanglement' hypothesis.

Consistency (Round 3): When predicting the likelihood of cumulatively consistent labels in Round 3 there was a significant main effect of experimental condition (β = -1.939, SE = 0.65, p < .005). It was more likely that established conventions would be broken by Round 3 for the Preconceived condition. There was also a significant (negative) main effect of animacy (β = -1.596, SE = 0.62, p < .05). It was more likely an established convention was *broken* rather than retained if a Round 3 label was animate. There was again a significant main effect of consistency with post-interaction labels produced later by Player A (β = 1.59, SE = 0.73, p < .05), so it was more likely a Round 3 label retained the established convention if it matched the post-interaction label produced by Player A (just like Round 2 Game III). And there was again no significant main effect of consistency with post-interaction labels by Player B or by the Learner, and once again a significant interaction effect of consistency with both at once (β = 2.724, SE = 1.33, p < .05). These results support our hypotheses on consistency with the exception of animacy which we expected would *improve* rather than hurt the likelihood of game-wide conventions. The pattern of effects from post-interaction label consistency repeats from Round 2 Game III – so that consistency with post-interaction labels by Player A or *both* Player B and the Learner aids convention.

Alignment: Finally, we consider our hypothesis linking within-game and post-interaction alignment. As depicted in **figure 8** the proportion of items for which post-interaction player labels were aligned is not as high in the Preconceived condition (53% vs. 68%): a pattern consistent with our hypothesis that any broken conventions in the matching game would result in misaligned post-interaction labels.

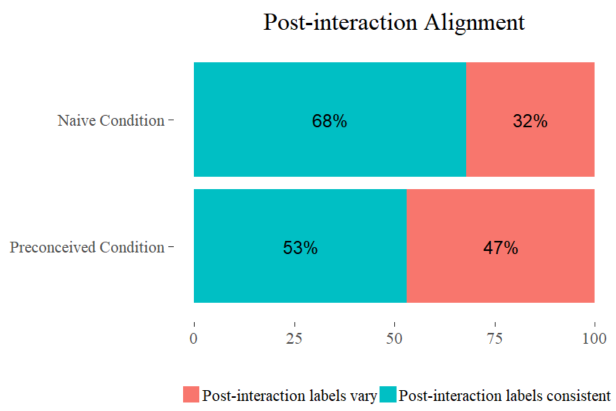


Figure 8. Alignment for post-game player labels across conditions.

To test this hypothesis, we constructed two mixed-effects logistic regression models in the same way described for consistency, to test if within-game alignment predicts post-interaction alignment. As discussed earlier, within-game alignment was represented by Round 3 cumulative consistency measures.

In our first model, we explored the effect of within-game alignment on post-interaction alignment across conditions. In our second model we explored how the pre-interaction labels produced by Preconceived Learners might have affected the above relationship.

When predicting the likelihood of post-interaction label alignment there was a significant main effect of within-game alignment (β = 4.187, SE = 0.53, p < .0001). It was much more likely post-interaction labels would be aligned, when within-game labels were also aligned (i.e. when participants only used a single convention). There was no significant main effect of condition though there was a significant interaction effect of condition and within-game alignment (β = -1.647, SE = 0.66, p < .05). The effect had an unexpected direction: it was *less* likely post-interaction labels would be aligned if within-game labels were also aligned for the Preconceived condition as opposed to the Naïve condition. This might connect to the qualitative ‘rogue label’ cases we considered over the previous section – inconsistency between game and post-interaction labels while a convention is established and maintained across the matching game. There was lastly a significant effect of consistency with rejected Matcher labels from Round 2: Game I on post-interaction alignment (β = 1.227, SE = 0.46, p < .01) though in the opposite direction than we had anticipated. Post-interaction label consistency with some rejected Matcher label *increased* the likelihood of post-interaction alignment unlike what we had hypothesised. There was no effect of label source, or animacy, on the likelihood a post-interaction label aligned with all the others.

Moving to the second model, when predicting the likelihood of post-interaction alignment just for the Preconceived condition, there was still a significant main effect of within-game alignment (β = 3.169, SE = 0.51, p < .0001) on post-interaction alignment. There was also still a significant effect of earlier rejected Matcher labels (β = 1.266, SE = 0.56, p < .05) in the same unexpected direction. There was a further significant main effect of consistency with any pre-interaction Learner labels (β = 3.473, SE = 0.89, p < .0005) but as with the case of rejected Matcher labels this was in the opposite direction than we had hypothesised: consistency with pre-interaction Learner labels made it hugely *more* likely that any post-interaction label would align with the other two rather than make post-interaction alignment less likely. This suggests the relationship between pre-interaction Learner labels and *post-interaction* alignment is not as antagonistic as we expected. (Recall that the presence of a pre-interaction Learner label curbed *within-game* alignment.) Adding further nuance to this relationship, though there was no significant main effect of label source, there was a significant interaction, between consistency with a pre-interaction Learner label and post-interaction labels produced *by the Learner* rather than Player B (β = -3.301, SE = 1.15, p < .005). This means that it was much less likely that a post-interaction label produced by the Learner would be aligned with the other two when the label was also consistent with the pre-interaction label for that item also by the Learner. This is consistent with our hypothesis that a pre-interaction Learner label would antagonise post-interaction alignment. It was also marginally non-significant that a post-interaction label by Player A rather than Player B matching any pre-interaction label by the Learner would make post-interaction label alignment less likely (β = -2.094, SE = 1.09, p = 0.052). Like with our previous model, there was no effect of animacy on post-interaction alignment.

5 Experiment 2 – Introduction

For our second set of results pairs of participants again played a collaborative matching game around two sets of cards, with players collaborating to place their individual cards in a mutually agreed order. And as before our aim was to disrupt otherwise-expected patterns of lexical alignment and refinement with (online) conceptual processing, elicited by presenting participants with an unexpected conceptual task. This time, however, the unexpected task was implicit – whereas in the first experiment they were told to conceptualise items – and it prompted another type of conceptual processing: feature inference.

The intent with our prompting implicit feature inference as part of an otherwise straightforward game by means of a conceptual-level inconsistency in an otherwise identical (and identically solvable) task, was to explore the extent to which conceptual processing, already shown to be active in dialogue for Experiment 1, was active even when implicitly required. And whether it would once again affect the lexical phenomena expected in this task, including lexical refinement and alignment, in a similar way.

As with our previous study the Director player established the target order, while the Matcher player arranged their own set of cards in that order to win the game. Unlike the previous study, the two sets of cards each belonging to the Director and Matcher were not identical versions of the same pictures: instead, they were two entirely different sets and consisted of written descriptions rather than images. Rather than matching the same *image*, participants tried to match two *descriptions of the same object* using a computer to check their answers, with the correct matches approved in the baseline condition and a predetermined incorrect/implausible set of matches approved by the software in the alternative.

Feedback given by the software administering the game made it possible to learn description pairings regardless of their plausible (in the baseline, where they did describe the same objects) or implausible (in the alternative) information content. Winning the game did not depend on processing descriptions or identifying any objects the card pairings were actually (or putatively, in the alternative) describing.

Regardless, we hypothesised that participants' implicit knowledge of physical objects would have an effect on the game by inviting the use of *de novo* conceptual labels where plausible, exchanging what words are on the cards for entirely new words identifying the objects like 'spoon' – vs. purely lexical refinement, where they are not. We expected that participants would revise labels into concept words for more efficiency when and where they can to expedite their coordination, and consequently that if such revision is not solicited by their background knowledge, their performance would suffer from it. Lastly, we hypothesised the use of concept words in the game would facilitate recall for their objects.

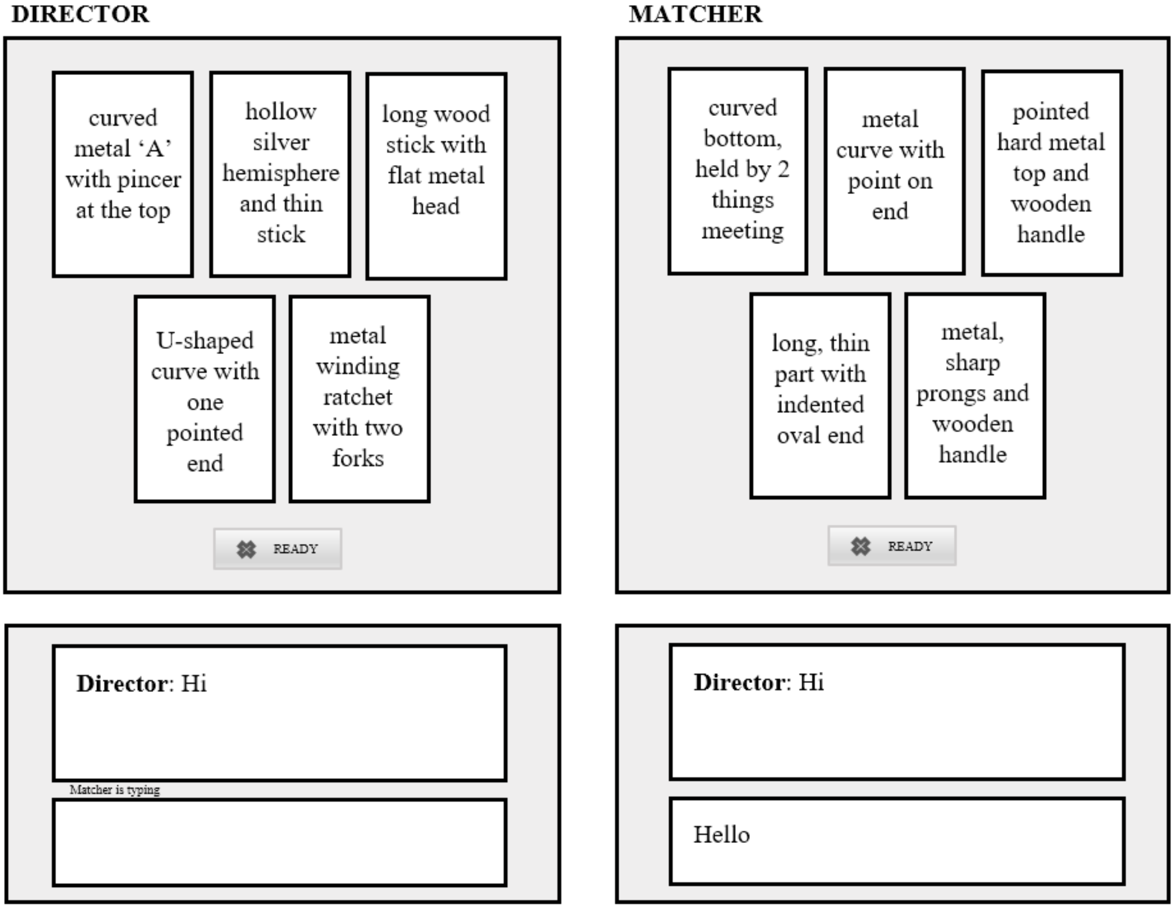
6 Method

i. Participants

Twenty pairs of native English speakers were recruited from the University of Edinburgh community and paid for their participation ($\bar{x}_{age} = 19$; 50% female). An additional nine pairs were recruited that failed to complete the task despite following instructions – more on this below – and three more pairs were excluded for being non-native speakers (two pairs), or not following task instructions (one pair). Participants in each pair were previously acquainted. Pairs were assigned to each condition randomly. In preparation for the experiment, fourteen native English speakers from the University of Edinburgh community were recruited on a voluntary/course credit basis, to provide the descriptions for the cards.

ii. Materials

The experiment was conducted in its entirety using a modified version of the Dialogue Experimental Toolkit (DiET; Mills & Healey, 2013). The items used to elicit the two sets of descriptions, which in turn were made available for participants to match during the experiment, are photographed museum objects from the online collection of the Metropolitan Museum of Art (metmuseum.org), available to use under the Creative Commons Zero License (CC0; creativecommons.org/publicdomain/zero/1.0/).



curved metal ‘A’ with pincer at the top

hollow silver hemisphere and thin stick

long wood stick with flat metal head

U-shaped curve with one pointed end

metal winding ratchet with two forks

Director: Hi

curved bottom, held by 2 things meeting

metal curve with point on end

pointed hard metal top and wooden handle

long, thin part with indented oval end

metal, sharp prongs and wooden handle

Director: Hi

Hello

Figure 1. Participant view of our modified chat client and description cards.

Five such photographs, of a hammer, a fishing hook, a spoon, a stirrup and a crossbow arming crank, were presented to the fourteen supplementary participants who were asked to describe each object in seven words or less, to help someone identify it based exclusively on its shape, texture and materials. From these fourteen descriptions, a pair of two descriptions per object was selected for the game and presented on screen as cards so that a complete set of cards contained one description for each object. The two sets of cards in the game thus contained each of two descriptions for those same five objects.

7 Results/Discussion

i. Variables/Coding

Variables were measured at both the game and individual item (e.g. hammer, hook) level. Game-level variables are *game length in minutes and rounds* and *test phase length in rounds*. Item-level variables, measured for each item (out of five per game), rather than each group (out of twenty), include *lexical vs. conceptual label* (initial and final), *recall*, *refinement*, *repeat*, *alignment* and *post-game alignment*.

The item-level binary variables merit more discussion. For the 'lexical vs. conceptual label' variable, we coded as 1 if a label included *at least one word not found in either description* for the appropriate matching pair (i.e. whatever pairing was correct for the item in that condition – cf. table 1); else as 0. A single exception was made for the label 'fork' for item #5 ('metal winding ratchet with two forks') when it was clear that label was intended as a noun (viz. *the* fork) rather than just a lexical shorthand. In the 'final' cases we only considered the last time the item was referred to, usually in the test phase, coded separately for Director and Matcher utterances. In the 'initial' cases we considered every label present for that item and coded as 1 if a rejected label met the above criterion – again for each player.

For the post-game 'recall' variable, we coded as 1 if a post-game description of that item was present whether or not that description or label was accurate or plausible, separately for each player, else as 0. 'Refinement' measured whether the final label used by each player was four or fewer words in length. 'Repeat' measured whether a post-game label had also been *produced during the game* by any player. Lastly, 'alignment' measured whether the final label per item for Director and Matcher was identical; 'post-game alignment' measured whether post-game labels for an item, if both present, were identical.

ii. Game Length and Label Type Frequency

Game: Game length in minutes and game length in rounds were both significantly longer for pairs in the implausible condition. Game length in minutes is almost doubled on average for implausible pairs ($\bar{x}_{tBase} = 28.01$ min, SD = 15.48, vs. $\bar{x}_{tExp} = 51.82$ min, SD = 15.62; t(18) = -3.42, p < .005) while game length in rounds is near triple ($\bar{x}_{RndBase} = 9$, SD = 6.58, vs. $\bar{x}_{RndExp} = 25.4$, SD = 12.62; t(18) = -3.64, p < .0001). When counting just rounds in the test phase, the difference from $\bar{x}_{TRndBase} = 4.2$ (SD = 5.32) to $\bar{x}_{TRndExp} = 7.5$ (SD = 9.74) was vastly reduced, and it was not statistically significant.

These results for game-level variables suggest our experimental manipulation has had great impact on overall task efficiency with implausible pairs slower to finish the task and requiring more feedback on average to learn correct pairings. Once those pairings are learned, the difference in efficiency between conditions is effectively eliminated. At the same time, it is useful to note how variable game length in rounds was between pairs (given its SD) potentially implying some participant pairs made much more efficient use of the feedback than others, particularly in the implausible condition. This could relate to why just under half of groups originally recruited in the implausible condition failed to finish the task; as opposed to no failed plausible condition groups, where the overall task efficiency was much higher.

Item: As with task efficiency our manipulation had obvious impact on the primary variable of interest at the item level, namely lexical vs. conceptual labels. This effect was particularly pronounced in final labels for the implausible condition, where *there were no final conceptual labels* for either Director or Matcher players, contrasted to a 82%/18% divide between lexical/conceptual final labels for plausible pairs, for both Director and Matcher players. To facilitate our intended analysis, we therefore chose to combine our measure of ‘initial’ and ‘final’ lexical vs. conceptual labels into a joint measure of where *any* conceptual label was attached to an item, whether that label was accepted as final, or later revised.

Results for our combined lexical-vs.-any-conceptual measure, for each player separately and for both players together, are summarised in Figure 2 by condition. The stark reversal in trend from individual to combined measure in the plausible condition, where the majority of labels were lexical when taken individually but conceptual when taken together, reflects the fact some items were given a conceptual label by only one player. When combining results for both players more items were labelled by either player at some point in the game (our combined conceptual vs. lexical measure) using concept words.

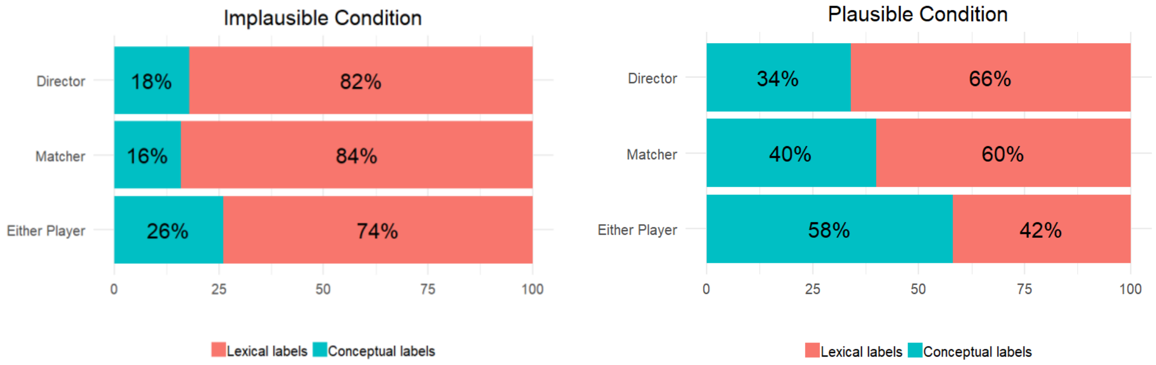


Figure 2. Relative frequency of lexical vs. conceptual label usage across conditions.

The lexical vs. conceptual measure is also summarised for specific items across both players and both conditions together in Figure 3. Item 5 (the crank) was given conceptual labels rather less often (25% total) than the other four (35-55%) consistent with its unintuitive nature. Proportions – per player and jointly – of recalled and refined labels are summarised by condition in Figure 4. Proportions of labels repeated post-game are explored in Figure 5. Finally, Figure 6 shows alignment for in-game and post-game labels. This last set of results suggests alignment in general was very limited in either condition.

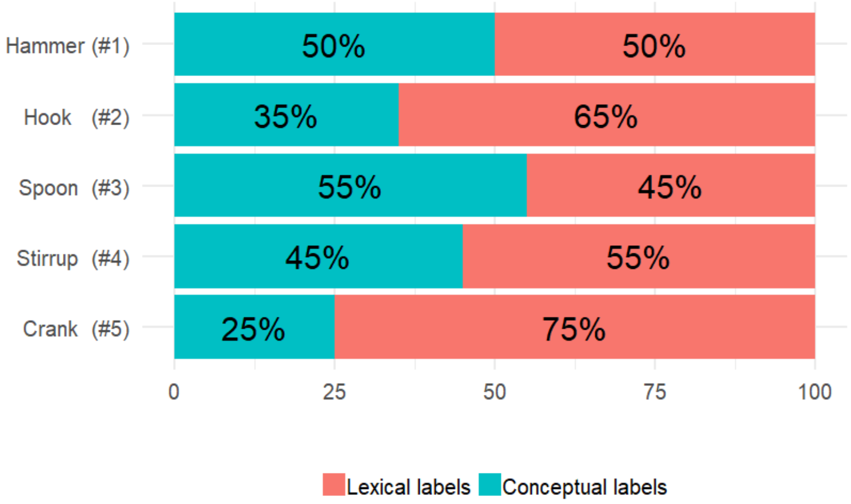


Figure 3. Relative frequency of lexical vs. conceptual label usage by individual item.

These results broadly support our experimental hypotheses, with caveats for efficiency and alignment.

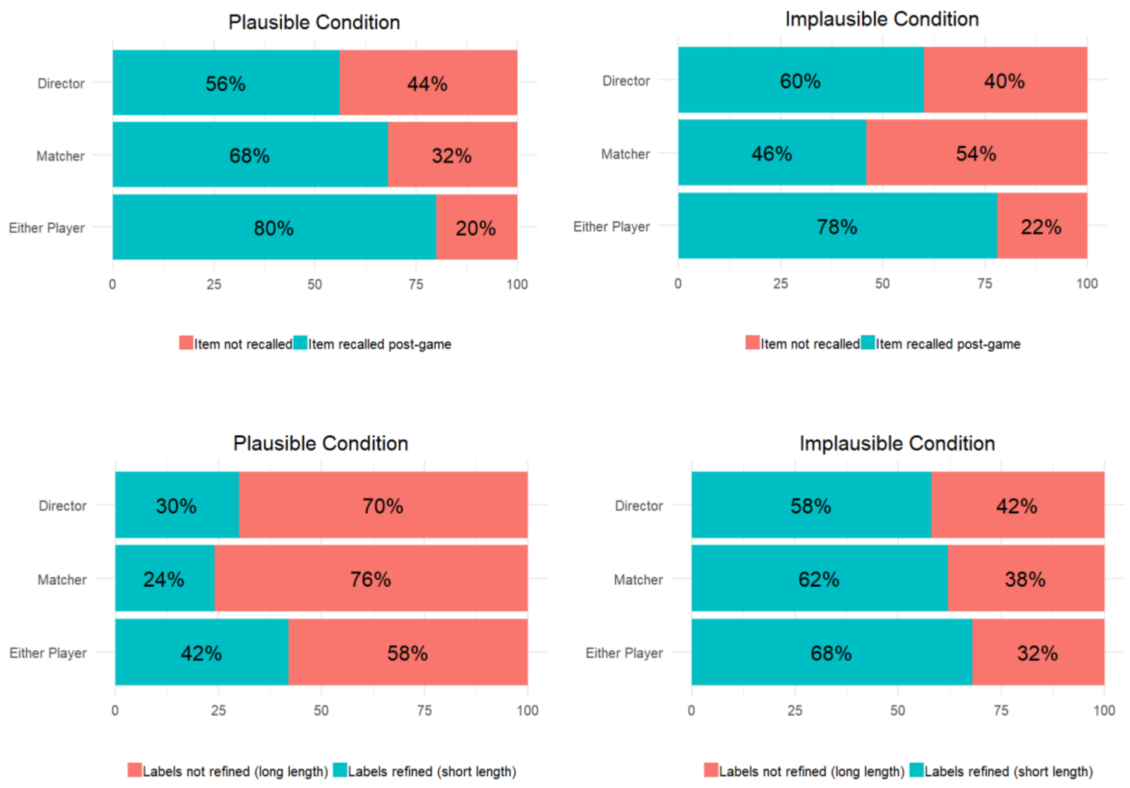


Figure 4. Relative frequency of post-game recall and label refinement across conditions.

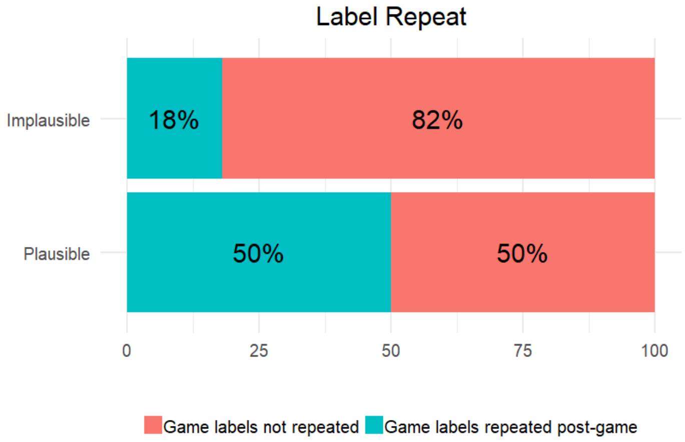


Figure 5. Relative frequency of labels repeated post-game across conditions.

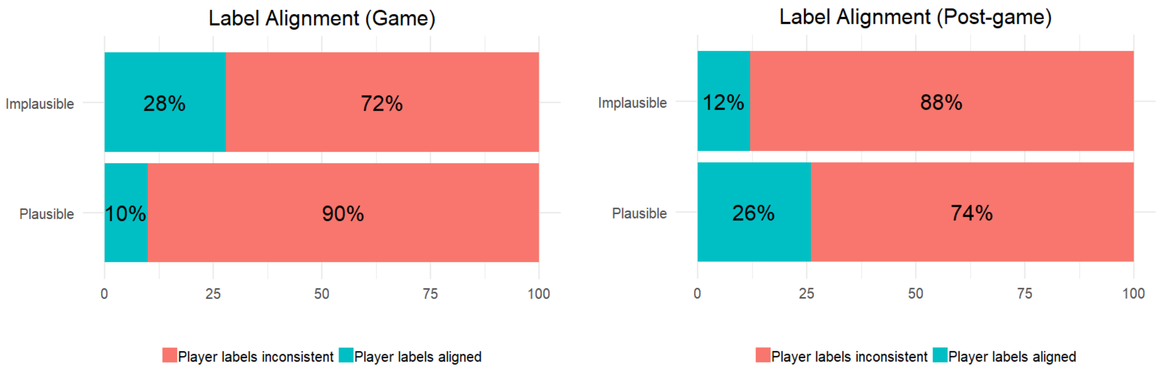


Figure 6. Relative frequency of aligned vs. inconsistent game and post-game labels across conditions.

iii. Logistical Models

Lexical vs. Conceptual: For this and the other item-level outcome variables we followed a consistent procedure, constructing a multilevel logistic regression model with participant pair added as a random intercept to predict the relevant variable: in this case, whether any label used for an item is conceptual. For each model we inserted predictors in order of theoretical importance (their order of mention here) retaining significant predictors that also significantly improved model fit, as we discuss above in §3iii.

As our primary hypothesis was focused on the use of conceptual labels in general rather than in one or another role, we omit models of individual player production. Instead, we aimed to predict how likely items were to be baptised by a conceptual label at least once per game whether by one or both players.

When predicting lexical vs. conceptual likelihood for item labels there was a significant main effect of condition (β =-3.17, SE = 1.16, p < .01), Director refinement (β = 2.28, SE = 0.9, p < .05) and the item being labelled: relative to the harder-to-conceptualise crank the hammer (β = 2.09, SE = 1.04, p < .05) and spoon (β = 2.29, SE = 1.07, p < .05), both significantly raised the likelihood of a conceptual label. There was no effect of Matcher refinement or of in-game alignment – nor any significant interactions.

Recall: Toward predicting whether items are likely to be recalled post-game, in line with the Director vs. Matcher difference in the effect of refinement on labels, we opted to model each player separately and to also divide the 'lexical vs. conceptual labels' variable accordingly when inserted as a predictor.

When predicting the likelihood of post-game item recall for the Director there was a significant effect of conceptual label use by the Director (β = 2.03, SE = 0.75, p < .01). There was no significant effect of conceptual label use by the Matcher, or of condition, refinement or of in-game alignment of labels.

When predicting the likelihood of post-game item recall for the Matcher there was a significant effect of conceptual label use by the Matcher (β = 2.00, SE = 0.77, p < .01). Yet there was also a significant effect of conceptual label use by the Director (β = 2.38, SE = 0.88, p < .01) and significant interaction between the two variables (β = -2.76, SE = 1.30, p < .05). The negative interaction coefficient implies that when Director and Matcher both give a conceptual label, the effect of the second conceptual label is negligible in light of the interaction. The *presence or absence* of a concept is making the difference. There was no significant effect of condition, refinement or in-game alignment on Matcher label recall.

Alignment: When predicting the likelihood of in-game label alignment, there was a significant main effect of Director refinement (β = 2.29, SE = 0.96, p < .05) and of Matcher refinement (β = 2.31, SE = 0.93, p < .05) but their positive effect on alignment was additive – there was no significant interaction. There was also no significant effect of lexical vs. conceptual labels, nor of the experimental condition.

Conversely when predicting the likelihood of post-game alignment there was a significant main effect of label repeat (β = 2.42, SE = 0.89, p < .01) – i.e. of whether a post-game label came from the game – and of conceptual label use by the Director (β = 1.596, SE = 0.72, p < .05). Yet there was no effect of conceptual label use by the Matcher, of refinement by either player, or of the experimental condition. Most unexpected of all, there was no significant effect of in-game alignment on post-game alignment.

8 General Discussion

Collectively, the above results showcase the presence of online conceptual processing in dialogue and their striking effects on alignment and other expected lexical processing behaviours. In so far as these conceptual behaviours can be interpreted pending further analysis, they seem consistent with the idea, that human interlocutors implicitly process conceptual information from lexical labels, both where the task demands are explicitly conceptual or descriptive (as with Experiment 1) and where no conceptual understanding of the problem is required to solve it (as with the supervised learning in Experiment 2).

Moreover, participants appear to actively juggle conceptualisations from different sources, preferring salient labels over less attractive alternatives, at the cost of their performance in the collaborative task. This was true when these labels break with established convention in Round 3 of Experiment 1 – and also where the labels participants had available were inconsistent in Experiment 2, where participants starkly sacrificed task performance to search for alternative labels, to replace unsuitable inferred ones.

Though these results are far from conclusive, even in their preliminary presentation across this chapter they argue for the broader view that conceptual coordination (and thereby coordination of information content for reference, which these concepts provide) could be given an empirical account based on the transmission of successful labels, and rejection of unsuccessful labels; not unlike the natural scientists envisioned by Frege in his own definition of how coordination could be achieved. Elaborated through further results, and further analysis of the above results, such an account is itself an initial elaboration of how conceptual content passes or is blocked from passing from one user to another over the course of normal communication. It thereby sheds light on conceptual effects on sociolinguistic transmission, and in particular on the ‘messy clash of perspectives’ that I previously suggested the received account of sociolinguistic transmission – even from studies of purely lexical conventions in psycholinguistics, such as those motivating the hypotheses overturned in Experiment 1 – had previously left unexplored.

Chapter 7: Point of Departure

1 Three Easy Pieces

> "[Neurotics] are people who build castles in the air, psychotics are people who live in them, while psychiatrists are people who charge the rent."
>
> (Rob Buckman, quoted by Richard Bentall in 'Madness Explained', 2003, p. 109.)

Throughout the above I pursued three overall objectives, which I will now recapitulate in turn.

As part of the initial four chapters I analysed three pre-eminent theories of reference with an emphasis on understanding where they each had most notably succeeded, and also where each had most notably failed. I framed this analysis from the viewpoint of three interrelated problems: contact, content, and coordination – which I related to each of the three theories, as achievements and overall perspectives.

The first theory, by Frege, excelled in describing reference as a logico-mathematical phenomenon and emphasised generality above all. To produce his theory Frege began with the question of coordinating knowledge, and developed the assertion that coordination is only possible under a shared language; an assertion I captured under the guise of commensurable specification, or rule CS. Having assured CS in his system of reference Frege offered a vision of reference as a tool for collaboration and coordination – one that I argued was more pragmatic than the traditional reading of his methodology might suggest. Yet though he captured coordination, and arguably the whole of reference in his limited mathematical context of ideal usage, Frege still failed to capture important cases of reference from the actual world. His system assumed that each label represented in it came with a referent, which is often not the case.

Instead, it fell to Russell to extend the locutions Frege gave for logic and reference, into a system that may capture the information content in every sort of label, empty or otherwise. Descriptivism carried with it the empiricist background Russell brought with him, much as Frege's system had carried logic. And I argued extensively for the essential character of this empiricist foundation, to Russell's method. Nonetheless, Russell too faced a limitation, in the form of implementing descriptivism (as with Frege) not as a logical exercise but in the actual world of messy, partial utterances, and underspecified labels. His empiricist descriptions were sufficient to describe empty names, but not sufficient to predict them.

In solving Russell's problem of how and when a label would successfully designate a referent, Kripke in turn brought his own preoccupations to the successful 'causal theory of reference'. Most notably he omitted (as I argued) a deeper account for how the causal links required for his theory are themselves established, and for how they are transmitted – so that *different* questions now remained unanswered.

The theories each roughly emphasised one aspect of reference, described a second and omitted a third – and I used the pattern of emphasis and omission to chart the gaps they opened relative to each other.

2 War of the Worlds

In setting out the philosophical background of these three major systems, my second objective was to explore their relation to contemporary scientific work on conceptual and perceptual processing. More specifically, I considered the link between the descriptive empiricism of Russell (with its emphasis on information content), and the contemporary psychological view of concepts as cognitive descriptions.

Taking descriptivism on one side and conceptual description on the other, I considered the many ways transplanting the Russellian descriptivist project from sense-data to conceptual description of referents can help provide a powerful definition for the content of reference, if concepts are that content. And in particular, I considered their description in the form of a generative graphical model aiming to capture their real properties and interrelations. With its capacity to represent context and efficiently limit what is important to include in describing a referent, I argued that this system is ideally placed to update the Russellian account of content into a more powerful (and far more flexible) 'conceptual descriptivism'.

Along similar lines, taking the Kripkean 'causal' criterion of contact on one side, and the most robust cognitive example of contact, *perception*, on the other, I explored how exactly perception achieves its accurate reconstruction of distal entities. Throughout the chapter-length discussion of how perception succeeds, I considered its flexible usage of assumptions, its consistent use of descriptions, and the fact that perception is so often right that perceptual processing proceeds on the basis that it is *always* right. I accordingly identified the criterion ensuring perceptual contact not in the truth of its descriptions but in the robustness of its mechanism – so that statistically speaking the descriptions are exactly and only of *what there is* much more often than they are not. I labelled this the 'statistical criterion' of contract; and I suggested the alternative by Kripke, the 'causal criterion' collapses into that statistical criterion.

Finally, as my third and primary objective, side-stepping coordination, I assembled the above two into a putative framework for reference combining conceptual descriptivism and the statistical criterion for contact. Making this possible is the analytical structure of a 'notional world' – the result of taking e.g. a conceptual description and using it to assert whatever that concept (as described) ought to satisfy. If repeated for each such belief about concepts, where these beliefs are presented as a generative model, then the notional world will simulate some portion of the actual world. I called this construct a World – and I gave a series of context-driven criteria for where a World would simulate some portion of the actual world stable enough over time that World would increasingly satisfy the statistical criterion of contact for that context. As a result, for each labelled concept in World, its presence in the construct alone is sufficient to meet the statistical criterion of contact by substitution, and as it *is a description* itself it will also trivially meet the criterion of content by conceptual description. As a result, I argued that the problem of (at least) contact and coordination can be considered solved for suitable contexts, just in having a conceptual description of objects in these contexts that is sufficiently adapted to them. (Of which a corresponding World is just the analytical construct demonstrating that this is a solution.)

Therefore Worlds are castles in the air – but under the right circumstances they can be inhabited, and for those circumstances (themselves equivalent to the ideal language described by Frege at the outset) the problem of reference can at least be two-thirds solved. The remaining question of coordination is the purview of the last chapter, where I briefly summarise two sets of experimental results over how conceptual processing interacts with, and even overcomes other language processing. In showcasing the ever-present role of concepts, and considerations such as the *quality* and *alignment* of concepts in regular natural language dialogue, these results establish a foothold for tackling the full mechanics of a scientific theory of conceptual information in language and dialogue – and conceptual coordination.

References

Andrewes, D. (2001). *Neuropsychology: From theory to practice*. Hove: Psychology Press.

Arsdall, J. E., Nairne, J. S., Pandeirada, J. N., & Cogdill, M. (2015). Adaptive memory: Animacy effects persist in paired-associate learning. *Memory, 23*(5), 657-663.

Ashby, F. G. & Maddox, W. T. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, *53*(1), 49-70.

Bálint, R. (1909). Seelenlähmung des 'Schauens,' optische Ataxie, räumliche Störung der Aufmerksamkeit. *Monatschr. Psychiat. Neurol. 25,* 51-81

Barlow, H. B. (1953). Summation and inhibition in the frog's retina. *Journal of Physiology*, *119*(1), 69-88.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3).

Barsalou, L. W. (1987) The instability of graded structure: implications for the nature of concepts. In U. Neisser (Ed.), *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization* (pp 101-140). Cambridge: Cambridge University Press.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Beaney, M. (1996). *Frege: Making Sense.* Bristol: Bristol Classical Press.

Beaney, M. (1997). *The Frege Reader*. Oxford: Blackwell.

Benacerraf, P., & Putnam, H. (1983). *Philosophy of mathematics: Selected readings* (Second ed.). Cambridge: Cambridge University Press.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* Springer.

Bonin, P. (2004). *Mental Lexicon: Some Words to Talk about Words.* Nova Publishers.

Boring, E. G. (1930). A new ambiguous figure. *American Journal of Psychology, 42,* 444-445.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(6), 1482-1493.

Broca, P. (1865). Sur le siege de la faculte du langage articule. *Bulletin de la Societe d'anthropologie, 6,* 337-393.

Burge, T. (1973). Reference and proper names. *Journal of Philosophy, 70*(14), 425-439.

Burge, T. (1979). Individualism and the Mental. *Midwest Studies In Philosophy*, *4*, 73-121.

Caldwell, C. A., & Millen, A. E. (2008). Studying cumulative cultural evolution in the laboratory. *Philosophical Transactions of the Royal Society. Series B: Biological Sciences, 363*(1509), 3529–3539.

Caldwell, C. A., & Smith, K. (2012). Cultural evolution and perpetuation of arbitrary communicative conventions in experimental microsocieties. *PLoS ONE 7*(8): e43807.

Carruthers, P. (1984). Eternal Thoughts. *The Philosophical Quarterly, 34*(136), 186-204.

Casasanto, D. & Lupyan, G. (2015). All concepts are ad hoc concepts. In E. Margolis & S. Laurence (Eds.), *Concepts: New Directions.* Cambridge, MA: MIT Press.

Chalk, M., Seitz, A. R., & Seriès, P. (2010). Rapidly learned stimulus expectations alter perception of motion. *Journal of Vision, 10*(8).

Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge, MA: MIT Press.

Clark, A. (1992). Presence of a symbol. *Connection Science, 4*(3), 193-205.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioural Brain Sciences, 36*(3), 181-204.

Clark, A. (2015). *Embodied Prediction.* In T. Metzinger & J. M. Windt (Eds), Open MIND: 7(T). Frankfurt am Main: MIND Group.

Clark, A. (2016). *Surfing Uncertainty: Predication, Action, and the Embodied Mind.* New York: Oxford University Press.

Clark, H. H., & Marshall, C. (1981). Definite knowledge and mutual knowledge. In A. Joshi, Bruce H. Weber & Ivan A. Sag (Eds.), *Elements of Discourse Understanding.* Cambridge: Cambridge University Press.

Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science, 13*(2).

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22,* 1-39.

Clark, H. H., & Wilkes-Gibbs, D. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language, 31,* 183-194.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49*(12), 997-1003.

Corlett, P. R., Frith, C. D., & Fletcher, P. C. (2009). From drugs to deprivation: A Bayesian framework for understanding models of psychosis. *Psychopharmacology, 206*(4), 515-530.

Crutch, S. J., & Warrington, E. K. (2008). The influence of refractoriness upon comprehension of non-verbal auditory stimuli. *Neurocase,* 14, 494-507.

Cutting, J. E. (1991). Why our stimuli look as they do. In G. Lockhead & J.R. Pomerantz (Eds.), *The perception of structure: Essays in honor of Wendell R. Garner* (pp. 41-52). Washington, DC: American Psychological Association.

Dam, W. O. van, & Hommel, B. (2010). How object-specific are object files? Evidence for integration by location. *Journal of Experimental Psychology: Human Perception and Performance, 36,* 1184-1192.

Danks, D. (2014). Perception, causation, and objectivity. *Mind, 123*(490), 635-639.

Descartes, R. (1996). *Meditations on first philosophy.* (J. Cottingham, Trans.). Cambridge: Cambridge University Press. (Original work published 1641).

Diamond, C. (1984). What does a concept script do? *The Philosophical Quarterly, 34*(136), 343-368.

Druks, J., & Shallice, T. (2000). Selective preservation of naming from description and the ‘restricted preverbal message’. *Brain and Language, 72*(2), 100-128.

Donnellan, K. (1966). Reference and definite descriptions.

Dummett, M. (1973). *Frege: Philosophy of language*. London: Duckworth.

Dummett, M. (1976). Frege as a Realist. *Inquiry, (19)*, 455-468

Dummett, M. (1993). *Origins of Analytical Philosophy.* Cambridge, Massachusetts: Harvard University Press.

Ellis, W. D. (1938). *A Source Book of Gestalt Psychology.* New York: Harcourt, Brace & World.

Enderton, H. (2001). *A mathematical introduction to logic* (Second ed.). London: Academic Press.

Evans, G. (1982). *The Varieties of Reference.* Oxford: Clarendon Press.

Fisher, A. V., Godwin, K. E., Matlen, B. J., & Unger, L. (2015). Development of category-based induction and semantic knowledge. *Child Development, 86,* 48-62.

Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience, 10,* 48-58.

Fodor, J. A. (1975). *The Language of Thought*. Harvard: Harvard University Press.

Fodor, J. A. (2010). *The Language of Thought Revisited*. Oxford: Oxford University Press.

Frege, G. (1879). *Begriffsschrift. Eine der arithmetischen nachgebildete Formelsprache des reinen Denkens.*

Frege, G. (1884). *Grundlagen der Arithmetik. Eine logisch-mathematische Untersuchung über den Begriff der Zahl*. Breslau: Verlage Wilhelm Koebner.

Frege, G. (1891). *Funktion und Begriff.* Jena: Herman Pohle.

Frege, G. (1892a). Über Sinn und Bedeutung. *Zeitschrift für Philosophie Und Philosophische Kritik, 100*, 25-50.

Frege, G. (1892b). Über Begriff und Gegenstand. *Vierteljahrsschrift fir wissenschaftliche Philosophie, 16,* 192-205.

Frege, G. (1894). Rezension von: E. Husserl, Philosophie der Arithmetik I. In *Zeitschrift für Philosophie und philosophische Kritik*, *103*, 313-332.

Frege, G. (1897). Logik. In M. Beaney (Ed.), *The Frege Reader*. Oxford: Blackwell.

Frege, G. (1904). Was ist eine Funktion? In S. Meyer (Ed.), *Festschrift Ludwig Boltzmann gewidmet zum sechzigsten Geburtstage* (656-666). Leipzig: Barth.

Frege, G. (1914). *Logik in der Mathematik*. Unpublished manuscript, University of Jena. In G. Gabriel (Ed.), Vorlesungen über Begriffsschrift, nach der Mitschrift von Rudolf Carnap. *History and Philosophy of Logic, 17,* 1–48.

Frege, G. (1918). The Thought: A logical inquiry. In Beaney, M. (Ed.), *The Frege Reader*. Oxford: Blackwell.

Frege, G. (1919). Notes for Ludwig Darmstaedter. In Beaney, M. (Ed.), *The Frege Reader*. Oxford: Blackwell.

Frege, G. (1953). *The Foundations of Arithmetic: A Logico-Mathematical Enquiry into the Concept of Number.* (J. L. Austin, Trans.). New York, NY: Harper. (Original work published 1884).

Frege, G. (1956). The Thought: A logical inquiry. (P. T. Geach, Trans.). In *Translations from the philosophical writing of Gottlob Frege.* Oxford: Oxford University Press.

Frege, G. (1967). Begriffsschrift, a formula language, modeled upon that of arithmetic, for pure thought. (S. Bauer-Mengelberg, Trans.). In Heijenoort, J. van (ed.), *From Frege to Gödel, a source book in mathematical logic*. Cambridge, MA: Harvard University Press. (Original world published 1879).

Frege, G. (1997). Function and Concept. In D. H. Mellor & Alex Oliver (eds.), *Properties*. Oxford University Press. 130-149.

Fridriksson, J., Fillmore, P., Guo, D., & Rorden, C. (2015). Chronic Broca’s aphasia is caused by damage to Broca’s and Wenicke’s areas. *Cereb Cortex, 25*(12), 4689-4696.

Frith, C. (2005). The neural basis of hallucinations and delusions. *Comptes Rendus Biologies, 328*(2), 169-175.

Frith, C. (2007). The social brain? *Philosophical Transactions of the Royal Society of London. B: Biological Sciences, 362*(1480), 671-678.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transation of the Royal Society of London: B, Biological Sciences, 360*(1456), 815-836.

Fussell, S. R., & Krauss, R. M. (1989). The effects of intended audience on message production and comprehension: Reference in a common ground framework. *Journal of Experimental Social Psychology, 25,* 203-219.

Garrod, S., & Doherty, G. (1994). Conversation, co-ordination and convention: An empirical investigation of how groups establish linguistic conventions. *Cognition, 53,* 181-215.

Garrod, S., Fay, N., Lee, J., Oberlander, J., & Macleod, T. (2007). Foundations of representation: Where might graphical symbol systems come from? *Cognitive Science, 31*(6), 961-987.

Garrod, S., Fay, N., Rogers, S., Walker, B., & Swoboda, N. (2010). Can iterated learning explain the emergence of graphical symbols? *Interaction Studies, 11*(1), 33-50.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition, 23*(3), 183-209.

Gelman, S. A., & Davidson, N. S. (2013). Conceptual influences on category-based induction. *Cognitive Psychology, 66*(3), 327-353.

Gibson, J. J. (1950). *The perception of the visual world.* Boston: Houghton Mifflin.

Gibson, J. J. (1961). Ecological optics. *Vision research, 1,* 253-262.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston: Houghton Mifflin.

Gibson, E. J. (1987). Introductory Essay: What does infant perception tell us about theories of perception? *Journal of Experimental Psychology: Human Perception and Performance, 13*(4), 515-523.

Gillebert, C. R., & Humphreys, G. W. (2015). Mutual interplay between perceptual organization and attention: A neuropsychological perspective. In J. Wagemans (Ed.), *Oxford Handbook of Perceptual Organization.* Oxford: Oxford University Press.

Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin & Review, 9*(3), 558-565.

Godfrey-Smith, P. (2003). *Theory and reality: An introduction to the philosophy of scienc*e. Chicago: University of Chicago Press.

Gopnik, A., & Astington, J. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Development, 59*(1), 26-37.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: causal models, Bayesian learning mechanisms, and the ‘theory theory’. *Psychological Bulletin, 138*(6), 1085-1108.

Gowers, T. (2002). *Mathematics: A very short introduction*. Oxford: Oxford University Press.

Gregory, R. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 290*(1038), 181-197.

Hart, J., Sloan Berndt, R., & Caramazza, A. Category-specific naming deficit following cerebral infarction. *Nature, 316,* 439-440.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity mappings in deep residual networks. *arXiv preprint arXiv:1603.05027.*

Heck, R., & May, R. (2008). Frege's Contribution to Philosophy of Language. In *The Oxford Handbook of Philosophy of Language*.

Heck, R. (2015). Consistency and the theory of truth. *Review of Symbolic Logic, 8*(3), 424-466.

Heck, R., & May, R. (forthcoming). Truth in Frege. In M. Glanzberg (ed.), *Oxford Handbook of Truth*. Oxford: Oxford University Press.

Helmholtz, H. von. (1867). *Handbuch der physiologischen Optik.* Leipzig: Leopold Voss.

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Science, 11*(10), 428-434.

Hodges, J. R., Patterson, K., Oxbury, S., & Funnell, E. (1992). Semantic dementia: progressive fluent aphasia with temporal lobe atrophy. *Brain, 115*(6), 1783-1806.

Hoenig, K., Sim, E. J., Bochev, V., Herrnberger, B., & Kiefer, M. (2008). Conceptual flexibility in the human brain: dynamic recruitment of semantic maps from visual, motor, and motion-related areas. *Journal of Cognitive Neuroscience, 20*(10), 1799-1814.

Hohwy, J. (2013). *The Predictive Mind.* Oxford: Oxford University Press.

Hollingworth, A., & Rasmussen, I. P. (2010). Binding objects to locations: The relationship between object files and visual working memory. *Journal of Experimental Psychology: Human Perception and Performance, 36,* 543-564.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat’s striate cortex. *Journal of Physiology, 150,* 574-591.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex. *Journal of Physiology, 160*(1), 106-154.

Humphreys, G. W. (2016). Feature confirmation in object perception: Feature integration theory 26 years on from the Treisman Bartlett lecture. *Quarterly Journal of Experimental Psychology, 69*(10), 1910-1940.

Humphreys, G. W., & Riddoch, J. (2013). *A Case Study in Visual Agnosia Revisited: To See But Not To See.* Hove: Psychology Press.

James, W. (1895). The knowing of things together. *Psychological Review, 2*(2), 105-124.

Jee, B. D., & Wiley, J. (2014). Learning about the internal structure of categories through classification and feature inference. *The Quarterly Journal of Experimental Psychology, 67*(9), 1786-1807

Jeffrey, R. (2006). *Formal logic: Its scope and limits* (5th ed.). Indianapolis, Ind.: Hackett Pub.

Kahneman, D., & Treisman, A. (1984). Changing views of attention and automaticity. In R. Parasuraman & D.R. Davis (Eds.), *Varieties of Attention* (pp. 29-61). Orlando: Academic Press.

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The review of object files: Object-specific integration of information. *Cognitive Psychology, 24,* 175-219.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology, 55,* 271-304.

Kimmig, A., Mihalkova, L., & Getoor, L. (2015). Lifted graphical models: A survey. *Machine Learning, 99*(1), 1-45.

Kluge, E. W. (1980). Frege, Leibniz, and the notion of an ideal language. *Studia Leibnitiana, 12*(140).

Koller, D. & Friedman, N. (2009). *Probabilistic graphical models: principles and techniques.* Cambridge, MA: MIT Press.

Kornmeier, J., & Bach, M. (2006). Bistable perception: Along the processing chain from ambiguous visual input to a stable percept. *International Journal of Psychophysiology, 62,* 345-349.

Kornmeier, J., & Bach, M. (2009). Object perception: When our brain is impressed but we do not notice it. *Journal of Vision, 9,* 1-10.

Krauss, R. M., & Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction. *Psychonomic Science, 1,* 113-114.

Kripke, S. (1973). *Reference and existence: The John Locke lectures.* Oxford: Oxford University Press.

Kripke, S. (1980). *Naming and Necessity.* Massachusetts: Harvard University Press.

Kripke, S. (2013). *Reference and existence: The John Locke lectures.* Oxford: Oxford University Press.

Kuffler, S. W. (1953). Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology, 16*(1), 37-68.

Lambon Ralph, M. A., Jefferies E., Patterson K., Jones, R. W. (2009). Comprehension of concrete and abstract words in semantic dementia. *Neuropsychology, 23,* 492-499.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioural and Brain Sciences, 22,* 1-75.

Lewis, D. (1973). Causation. *Journal of Philosophy, 70*(17), 556-567.

Lichtheim, L. (1884). Über Aphasie. *Deutsches Archiv für klinische Medicin, 36,* 204-268

Lupyan, G. (2008). From chair to ‘chair’: A representational shift account of object labeling effects on memory. *Journal of Experimental Psychology: General, 137*(2), 348-369.

Lupyan, G. (2012). Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontier Psychology, 3*(54).

Lupyan, G. (2015). The centrality of language in human cognition. *Language Learning, 66*(3).

Lupyan, G., & Bergen, B. (2015). How language programs the mind. *Topics in Cognitive Science, 8*(2).

Lupyan, G., & Clark, A. (2016). Words and the world: Predictive coding and the language-perception-cognition interface. *Current Directions in Psychological Science, 24*(4), 279-284.

Lupyan, G., & Lewis, M. (2017). From words-as-mappings to words-as-cues: The role of language in semantic knowledge. *PsyArxiv Preprint.* doi:10.17605/OSF.IO/F83AU.

Lupyan, G., Rakison, D. H., McClelland, J. L. (2007). Language is not just for talking: redundant labels facilitate learning of novel categories. *Psychological Science, 18*(2), 1077-1083.

Lupyan, G., & Thompson-Schill, S. L. (2012). The evocative power of words: Activation of concepts by verbal and nonverbal means. Journal of Experimental Psychology: General, 141, 170-186.

Lupyan, G., & Ward, E. J. (2013). Language can boost otherwise unseen objects into visual awareness. *Proceedings of the National Academy of Sciences, 110*(35), 14196-14201.

MacKay, D. J. (2003). *Information theory, inference, and learning algorithms.* Cambridge: Cambridge University Press.

Macpherson, F. (2012). Cognitive penetration of colour experience: rethinking the issue in light of an indirect mechanism. *Philosophy and Phenomenological Research, 84*(1).

Mahon, B. Z., & Caramazza, A. (2009). Concepts and categories: a cognitive neuropsychological perspective. *Annual Review of Psychology, 60,* 27-51.

Maloney, L. T., Zhang, H. (2010). Decision-theoretic models of visual perception and action. *Vision Research, 50*(23), 2362-2374.

Malt, B. C. (1990). Features and beliefs in the mental represention of categories. *Journal of Memory and Language, 29*(3), 289-315.

Malt, B. C., & Sloman, S. A. (2007). Category essence or essentially pragmatic? Creator's intention in naming and what's really what. *Cognition, 105,* 615-648.

Marr, D. (1974). On the purpose of low-level vision. *MIT: AI Laboratory Memo 324.*

Marr, D. (1976). Early processing of visual information. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 275*(942), 483-519.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* New York: W. H. Freeman.

Marr, D., & Hildreth, E. (1980). Theory of edge detection. *Proceedings of the Royal Society of London. Series B, Biological Sciences, 207,* 187-217.

Marr, D., & Nishihara, H. K. (1976). *Representation and recognition of spatial organization of three-dimensional shapes.* A.I. Memo 377. Cambridge, Mass: MIT Artificial Intelligence Laboratory.

McDowell, J. (1984). De re senses. *The Philosophical Quarterly, 34*(136), 283-294.

McIntosh, R. (2018). Simple dissociations for a higher-powered neuropsychology. *Cortex.*

McNamara, T. P., & Altarriba, J. (1988). Depth of spreading activation revisited: Semantic mediated priming occurs in lexical decision. *Journal of Memory & Language, 27,* 545-559.

McNamara, T. P., & Healy, A. F. (1988). Semantic, phonological, and mediated priming in reading and lexical decisions. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 14,* 398-409.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*(3), 207-238.

Medin, D. J., & Shoben, E. J. (1988). Context and structure in conceptual combination. *Cognitive Psychology, 20,* 158-190.

Meinong, A. (1904). *Über Gegenstandstheorie.*

Mesoudi, A., & Whiten, A. (2008) The multiple roles of cultural transmission experiments in understanding human cultural evolution. *Philosophical Transactions of the Royal Society. Series B: Biological Sciences, 363,* 3489-3501.

Miceli, G., Capasso, R., Daniele, A., Esposito, T., Magarelli, M., & Tomaiuolo, F. (2000). Selective deficit for people's names following left temporal damage: An impairment of domain-specific conceptual knowledge. *Cognitive Neuropsychology, 17*(6), 489-516.

Miller, E. K. (1999). Straight from the top. *Nature, 401*(6754), 650-651.

Millikan, R. G. (1989). Biosemantics. *Journal of Philosophy, 86,* 281-297.

Millikan, R. G. (2004). *Varieties of Meaning.* Cambridge: MIT Press.

Mills, G. J. (2014). Dialogue in joint activity: Complementarity, convergence, conventionalization. *New Ideas in Psychology, 32,* 158-173.

Mills, G. J., & Healey, P. G. T. (2006). Participation, Precedence and Co-ordination in Dialogue. In *Proceedings of the 28th Annual Conference of the Science Society.* Vancouver, Canada.

Mills, G. J., & Healey, P. G. T. (2013). A dialogue experimentation toolkit.

Milne, P. (1986). Frege's Context Principle. *Mind, 95*(380), 491-495.

Mitroff, S. R., Scholl, B. J., & Noles, N. S. (2007). Object files can be purely episodic. *Perception, 36,* 1730-1735.

Mitroff, S. R., Scholl, B. J., & Wynn, K. (2005). The relationship between object files and conscious perception. *Cognition, 96,* 67-92.

Murphy, G. (2002). *The big book of concepts*. Cambridge, MA.; London: MIT Press.

Murphy, G. L., & Ross, B. H. (2010). Category vs. object knowledge in category-based induction. *Journal of Memory and Language, 63,* 1-17.

Newen, A., & Vetter, P. (2017). Why the cognitive penetration of our perceptual experience is still the most plausible account. *Consciousness and Cognition, 47,* 26-37.

Ng, A. Y., & Jordan, M. I. (2001). On discriminative vs. generative classifiers: A comparison of logistic regression and naïve Bayes. *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic,* 841-848.

Noonan, H. (1984). Fregean Thoughts. In Wright C. (Ed.), *Frege’s Conceptions of Numbers as Objects* (20-39). Aberdeen: Aberdeen University Press.

Orden, G. C. van, Pennington, B. F., & Stone, G. O. (2001). What do double dissociations prove? *Cognitive Science, 25*(1), 111-172.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97,* 185-200.

Palmer, S. E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition, 3*(5), 519-526.

Palmer, S. E. (1977). Hierarchical structure in perceptual representation. *Cognitive Psychology, 9,* 441-474.

Palmer, S. E. (1999). *Vision science: Photons to phenomenology.* Cambridge, MA; London: MIT Press.

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience, 8,* 976–987.

Peano, G. (1889). *Arithmetices principia, nova methodo exposita*. Rome.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference.* San Francisco, CA: Morgan Kaufmann.

Pearl, J. (2009). *Causality*. Cambridge: Cambridge University Press.

Pickering, M. J., & Garrod, S. (2004). Towards a mechanistic psychology of dialogue. *Behavioural and Brain Sciences, 27,* 169-190.

Pickering, M. J., & Garrod, S. (2013). Forward models and their implications for production, comprehension, and dialogue. *Behavioral and Brain Sciences, 36*(4), 377-392.

Pizlo, Z. (2001). Perception viewed as an inverse problem. *Vision Research, 41*(24), 3145-3161.

Poggio, T., & Ullman, S. (2013) Vision: Are models of object recognition catching up with the brain? *Annals of the New York Academy of Sciences, 1305*(1).

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*(3, Pt.1), 353-363.

Pothos, E., & Wills, A. (2011). *Formal approaches in categorization.* Cambridge: Cambridge University Press.

Putnam, H. (1975). The Meaning of ‘Meaning’. In H. Putnam (Ed.), *Philosophical Papers, Vol. II: Mind, Language, and Reality*. Cambridge: Cambridge University Press.

Pylyshyn, Z. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences, 22*(3), 341-365.

Quine, W. V. (1948). On What There Is. *Review of Metaphysics, 2*(5), 21-38.

R Core Team (2013). R: A language and environment for statistical computing. *R Foundation for Statistical Computing,* Vienna, Austria. URL http://www.R-project.org/.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Natural Neuroscience, 2*(1), 79-87.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulleting, 124*(3), 372-422.

Recanati, F. (2012). *Mental Files*. Oxford: Oxford University Press.

Reitter, D., & Lebiere, C. (2010). How groups develop a specialized domain vocabulary: A cognitive multi-agent model. *Cognitive Systems Research, 12,* 175-185.

Resnik, M. D. (1980). *Frege and the Philosophy of Mathematics.* Ithaca, NY: Cornell University Press.

Riddoch, M. J., & Humphreys, G. W. (1987). A case of integrative visual agnosia. *Brain, 110*(6), 1431-1462.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior, 14,* 665-681.

Rock, I. (1957). The effect of retinal and phenomenal orientation on the perception of form. *American Journal of Psychology, 70,* 493-511.

Rock, I. (1980). Difficulties with a direct theory of perception. *Behavioral and Brain Sciences, 3*(3), 398-399.

Rock, I. (1983). *The Logic of Perception.* Cambridge, MA: MIT Press.

Rock, I. (1997). *Indirect perception (MIT Press/Bradford Books series in cognitive psychology).* Cambridge, MA: MIT Press.

Rock, I., & Brosgole, L. (1964). Grouping based on phenomenal proximity. *Journal of Experimental Psychology, 67*(6), 531-538.

Rogers, T. T., & McClelland, J. L. (2004) *Semantic cognition: A parallel distributed processing approach.* Cambridge, MA: MIT Press.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology, 104*(3), 192-233.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models.* Cambridge, MA: MIT Press.

Russell, B. (1905). On Denoting. *Mind, 14*(56), 479-493.

Russell, B. (1906). On the Nature of Truth. *Proceedings of the Aristotelian Society, 7*(1), 28-49.

Russell, B. (1911). Knowledge by acquaintance and knowledge by description. *Proceedings of the Aristotelian Society, 11,* 108-128.

Russell, B. (1917). The relation of sense-data to physics. *Scientia, 16,* 1-27.

Sainsbury, R. (2005). *Reference without Referents*. Oxford: Clarendon Press.

Schulz, M. F., & Sanocki, T. (2003). Time course of perceptual grouping by color. *Psychological Science, 14,* 26-30.

Searle, J. R. (1958). Proper names. *Mind. 67*(266), 166-173.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences of the USA, 105*(15), 6424-6429.

Shallice, T. (1988). Information-processing models of consciousness: Possibilities and problems. In A. J. Marcel & E. Bisiach (Eds.), *Consciousness in Contemporary Science*. Oxford: Oxford University Press.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27,* 379-423.

Sloman, S. A., Love, B. C., & Ahn, W. K. (1998). Feature centrality and conceptual coherence. *Cognitive Science, 22,* 189-228.

Sloman, S. A. (2005). *Causal Models: How People Think About the World and Its Alternatives.* New York: Oxford University Press

Sloutsky, V. M. (2010). From perceptual categories to concepts: What develops? *Cognitive Science, 34*(7).

Sluga, H. (1980). *Gottlob Frege (Arguments of the philosophers).* London: Routledge and Kegan Paul.

Soames, S. (2007). The substance and significance of the dispute over two-dimensionalism. *Philosophical Books, 48,* 34-49.

Soames, S. (2010). *Philosophy of Language (Princeton foundations of contemporary philosophy)*. Princeton, New Jersey: Princeton University Press.

Strawson, P. F. (1959). *Individuals: An essay in descriptive metaphysics (University paperbacks)*. London: Methuen.

Strawson, P. F. (1964). Intention and convention in speech acts. *Philosophical Review, 73*(4), 439-460.

Szegedy, C., Vanhoucke, V., Ioffe, Sergey., Shlens, J., & Wojna, Z. (2015). Rethinking the inception architecture for computer vision. *arXiv preprint arXiv:1512.00567.*

Tarski, A. (1933). *Pojęcie prawdy w językach nauk dedukcyjnych*. Warsaw: Nakładem Towarzystwa Naukowego Warszawskiego.

Tarski, A. (1944). The Semantic Conception of Truth and the Foundations of Semantics. *Philosophy and Phenomenological Research, 4,* 341-376.

Teunisse, R. J., Cruysberg, J. R., Hoefnagels, W. H., Verbeek, A. L., & Zitman, F. G. (1996). Visual hallucinations in psychologically normal people: Charles Bonnet's syndrome. *Lancet, 347,* 794-797

Thelen, E., & Smith, L. B. (1994). *A dynamical systems approach to the development of perception and action.* Cambridge, MA: MIT Press.

Treisman, A. (1992). Perceiving and re-perceiving objects. *American Psychologist, 47*(7), 862-875.

Treisman, A. (1998). Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 353*(1373), 1295-1306.

Treisman, A. (1999). Solutions to the binding problem: Progress through controversy and convergence. *Neuron, 24,* 105-110.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97-136.

Treisman, A., & Gormican, S. J. (1988). Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review, 95*(1), 15-48.

Treisman, A. & Zhang, W. (2006). Location and binding in visual working memory. *Memory & Cognition, 34*(8), 1704-1719.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of Memory* (pp. 381-403). New York: Academic Press.

Twardowski, K. J. (1894). *Zur Lehre vom Inhalt und Gegenstand der Vorstellungen; Eine psychologische Untersuchung.* Vienna: Kluwer Academic Publishers.

Ullman, S. (1980). Against direct perception. *The Behavioural and Brain Sciences, 3*(3), 373-414.

Vetter, P., & Newen, A. (2014). Varieties of cognitive penetration in visual perception. *Consciousness and Cognition, 27,* 62-75.

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & Heydt, R. von der. (2012). A century of gestalt psychology in visual perception: Perceptual grouping and figure-ground organisation. *Psychological Bulletin, 138*(6), 1172-1217.

Warren, W. H. (2005). Direct perception: The view from here. *Philosophical Topics, 33,* 335-361.

Warren, W. H. (2013). Does computation theory solve the right problem? Marr, Gibson, and the goal of vision. *Perception, 41*(9), 1053-1060.

Warrington, E. K. (1975). The selective impairment of semantic memory. *Quarterly Journal of Experimental Psychology, 27*(4), 635-657.

Warrington, E. K., & McCarthy, R. (1983). Category specific access dysphasia. *Brain, 106*(4), 859-878.

Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain, 107,* 829-854.

Weiskopf, D. A. The plurality of concepts. *Synthese, 169*(145).

Wernicke, C. (1874). *Der Aphasische symptomencomplex*. Breslau: Cohn and Weigert.

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt. *Psycologische Forschund, 4,* 301-350.

Whitehead, A. N., & Russell, B. (1910, 1912, 1913). *Principia Mathematica* (Vols. 1-3). Cambridge: Cambridge University Press.

Wolfe, J. M., & Bennett, S. C. (1996). Preattentive object files: Shapeless bundles of basic features. *Vision Research, 37*(1), 25-43.

Wright, C. (1984). *Frege's Conceptions of Numbers as Objects.* Aberdeen: Aberdeen University Press.

Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition, 85*(3), 223-250.

Xu, F., Dewar, K., & Perfors, A. (2009). Induction, overhypotheses, and the shape bias: Some arguments and evidence for rational constructivism. In B. M. Hood & L. Santos (Eds.), *The Origins of Object Knowledge* (pp. 263-284.) Oxford: Oxford University Press.

Yee, E., Chrysikou, E. G., & Thompson-Schill, S. L. (2014). Semantic Memory. In Kevin Ochsner and Stephen Kosslyn (Eds), *The Oxford Handbook of Cognitive Neuroscience, Volume 1: Core Topics* (pp. 353-374). Oxford: Oxford University Press.

Yee, E., & Thompson-Schill, S.L. (2016). Putting concepts into context. *Psychonomic Bulletin and Review, 23,* 1015-1027.

Yee, E., Jones, M. N., & McRae, K. (2017). Semantic Memory. In J. T. Wixted & S. Thompson-Schill (Eds.), *The Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience (4th Edition, Volume 3: Language and Thought).* New York: Wiley.

Zalta, E. (1983). *Abstract objects: An introduction to axiomatic metaphysics*. Lancaster: Reidel.

Zwaan, R. A., & Taylor, L. J. (2006). Seeing, acting, understanding: motor resonance in language comprehension. *Journal of Experimental Psychology: General, 135*(1), 1-11.