THEORIES OF THE SELF: THE ROLE OF THE PHILOSOPHY AND NEUROSCIENCE OF LANGUAGE

WILLIAM STEPHEN JONES

THEORIES OF THE SELF: THE ROLE OF THE PHILOSOPHY AND NEUROSCIENCE OF LANGUAGE

A thesis submitted for the degree of

Doctor of Philosophy

by

William Stephen Jones

Department of Philosophy

University of Durham

2018

Supervisors:

Professor Wolfram Hinzen

Professor Hamish McAllister-Williams

Dr Andy Hanson

Dr Benedict Smith

Examiners:

Professor Andy Hamilton, University of Durham

Professor Markus Werning, Ruhr-University Bochum

Viva:

14th December 2018

DECLARATION

I confirm that no part of the material contained in this thesis has previously been submitted for any degree in this or any other university. All the material is the author's own work, except for quotations and paraphrases which have been suitably indicated. The copyright of this thesis rests with the author. No quotation from it should be published without his prior written consent, and information derived from it should be acknowledged.

William Stephen Jones

ACKNOWLEDGEMENTS

Firstly, I would like to express infinite gratitude to my primary supervisor, Professor Wolfram Hinzen. Wolfram has been a constant source of support for me during this PhD process. He has dedicated hundreds of hours to helping me develop as a researcher: listening to all of my half-baked ideas, motivating me through my intellectual lulls, commenting extensively on my writing, and advising me on practical aspects of life as a researcher. His vast knowledge, enthusiasm, ambition, patience, work ethic and dedication to his students has been an inspiration to me, both as a researcher and as a person. It is my hope that some of these traits will have seeped into my own psyche and will reverberate out from me as I proceed on to the next stage of my career.

My sincere thanks also goes to Professor Hamish McAllister-Williams, Dr Andy Hanson and Dr Benedict Smith who have played important roles in my supervisory team. Building on my MSc in Cognitive Neuroscience, Hamish and Andy advised me on the experimental aspects of this thesis, shepherding me through the dos and don'ts of EEG research, and the complex regulatory requirements of performing experimental work with NHS patients. Thanks also to Ben, for meticulously reading and commenting on the philosophical parts of the thesis, and for helping me to understand and organise the formal aspects of the PhD process.

I would like to thank my examiners, Professor Andy Hamilton and Professor Markus Werning, for their helpful comments on my thesis. Also, a special thanks to a wide range of other people who have helped me in a variety of different ways: Dr Txuss Martin and Dr Andrew Woodard, Dr Felicity Deamer, Professor Nicol Ferrier, Professor Douglas Turkington, Professor Rosemary Varley, Dr Stuart Watson, Dr Vitor Zimmer, Miss Helen Spencer, Dr Joy Allen, Dr Michael Power, Professor John Simpson, Dr Sara Graziadio, and several others…

Last but not the least, I would like to thank my family and friends for supporting me throughout this formative, and sometimes turbulent, period of my life… in particular, my fiancée and mother. Without their unwavering support, love and kindness, I doubt I would have ever achieved anything of worth in my life. Therefore, I dedicate this thesis to them.

ABSTRACT

The nature of self has been discussed for centuries, with myriad theories specifying propositions of the form 'The self is X'. Recently, psychology and neuroscience have added further such propositions and have sought to specify neural correlates for X. In this thesis, theories leading to all such propositions are subjected to methodological criticism. Specifically targeted are those theories that construct metaphysical, essentialist propositions on the nature of the self, and all other abstract concepts, more generally. On this point, it is concluded that theories of this type would benefit from taking into account the nature of language and the role it plays in the development of a theory. Theories that fail to consider language, run the risk of producing theoretical work affected by linguistic biases, those inherent to the language faculty. In this thesis, the biases of interest are referred to as 'noun phrase reification' and 'clausal reification', and an awareness of these is important for they can subvert the meanings of propositions, rendering them, and the theories built upon them, true by definition. In consideration of this methodological critique, a new theoretical approach to the self and to other abstract concepts is argued for. This new way of theorising combines aspects of basic scientific methodology with statistical modelling and linguistic theory, in which, the triangulation between two or more people, engaged in the act of naming objects and behaviours plays a formative role. Building on this new method in the context of the self, two experiments are performed using the electroencephalography neuroimaging technique; first in neurotypical adults, then in patients with a diagnosis of schizophrenia, a disorder commonly associated with an attenuated selfhood. It is hoped that this body of theoretical and empirical work will facilitate and catalyse progress on abstract concepts and their associated problems.

CONTENTS

PREFACE

The initial plan for this PhD research was to contribute to the traditional philosophical problems associated with the concept of self, examining and integrating aspects of psychology, cognitive neuroscience and schizophrenia research; the latter, because patients diagnosed with schizophrenia are often characterised as suffering from a diminished form of selfhood (Sass & Parnas, 2003; Sass, Pienkos, Nelson, & Medford, 2013).

This research began with a review of the self literature, but approximately halfway through this review an unexpected problem was identified. This was not one of the traditional problems associated with the self, but a methodological problem with how certain philosophical, psychological and neuroscientific theories approach the self, and other abstract concepts, more generally. This problem could not be bypassed, for it appeared to undermine a wide range of important theories. Consequently, the plan and focus of this thesis shifted, with significant attention now given to this problem: to articulating it, to exploring the reasons for its emergence, to computing the consequences of overlooking it, and to finding ways to solve/circumvent it, so that theorising on the self – and other abstract concepts – could proceed in a valid and efficient manner. Below is a succinct summary of the chapters, identifying the principal themes.

Chapter 1 reviews sixteen theories of the self, from philosophy, psychology and cognitive neuroscience. Based on this review, it is determined that when these individual theories of self are considered as independent theories, they often prove to be internally consistent and persuasive, but when they are considered from a birds-eye perspective, as a set of theories, significant contradictions arise, which undermines the current approach to the self and to other abstract concepts. Possible explanations and solutions to this state of affairs are considered, including the one-theory solution and the multiple (pattern theory) solution (Gallagher, 2013). The one theory solution argues that one particular theory of self is correct and all other contradictory theories are incorrect. The multiple theory solution argues that the plurality of incompatible theories is unproblematic, for a complex and jointly sufficient pattern of theories will account for the self, because the self is a cluster concept. Both positions are critically examined, and both are found wanting. The conclusions established in this chapter motivate a step-back from the frontline (first-order) debate of self (in Chapter 2), to consider more foundational and metatheoretical areas relating to nouns/concepts; ‘noun’ is the linguistic term for a class of words that name things. Concrete nouns name physical objects (e.g. ‘trousers’, ‘motorbikes’, ‘bananas’) and abstract nouns name ideas, states or qualities (e.g.

'politeness', 'justice', 'self'). Where possible, the term 'noun' is used in this thesis, as a synonym of 'concept', for the latter is theory-laden, whereas the former is relatively neutral.

Chapter 2 examines how the meanings associated with concrete and abstract nouns arise and function in humans. It is argued that they arise in different ways and are subject to different cognitive processes. Concrete nouns are typically acquired ostensively, through perceptuo-linguistic triangulation (PLT), where there is a three-fold, cross-modal, perceptual interaction between two or more individuals and an object in the world; whereas abstract nouns, devoid of a clear third-point physical object in the world, are acquired through point-to-point (P2P) interactions between two individuals, with one giving the information and one receiving the information. However, this P2P information can be supplemented by a form of triangulation involving two or more individuals and a reference to a behaviour, this is termed: 'perceptuo-linguistic triangulation of behaviour' (PLTb). This notion of PLTb is instrumental to the positive aspects of this thesis, developed in Chapter 4.

With a basic understanding of how concrete and abstract noun meanings arise, Chapter 2 then addresses their underlying meaning structure. This starts with a review and critique of the classical explanation of concept structure; where 'concept', is the historically-dominant term used to refer to information anchored to nouns. The classical theory conceives of concepts as rigidly definitional, comprising simpler properties that are individually necessary and jointly sufficient for concept membership. In this approach, these properties are explicated through deductive reasoning, called 'conceptual analysis'. Several problems with the classical theory are highlighted, leading to the conclusion that this is an inadequate approach to explicating concepts and solving their associated problems. Many theorists would probably be in (relative) agreement on this point, for an endorsement or acknowledgment of this form of analysis is rarely seen. However, in Chapter 3, it is demonstrated that much theoretical and empirical work – including the positions outlined in Chapter 1 – consciously or unconsciously assume and partake in aspects of this classical approach. Particularly, when they construct essentialist, metaphysical propositions of the form X is Y. Participation in this form of theorising has detrimental consequences to the validity of the research, also detailed in Chapter 3.

The critique of the classical theory is followed by a description and analysis of prototype theory, as an alternative, statistical, approach to explaining word/noun/concept structure. Prototype theory holds that word meanings have a probabilistic structuring, containing a set of properties (physical and functional) that are individually weighted in the word in question, each word specifying graded properties that its members tend to possess, a

numerical probability between 0 and 1. It is argued that this theory better explains the structuring of noun meanings, because it accounts for, and predicts, the empirical data, as well as satisfying important theoretical requirements. After establishing prototype theory as a more viable explanation of word structure, the nature and structuring of concrete and abstract nouns are described from within the prototypical perspective. It is concluded that concrete nouns have stable meanings, because the objects they name have a shared existence in the perceptual domain and are thus susceptible to PLT; PLT being the stabilising force. Whereas, abstract nouns are unstable, since they name things lacking a shared existence in the perceptual domain, and thus are not susceptible to PLT. The consequences of these two propositions play a significant role in this thesis and are discussed in detail in Chapter 4.

The next stage of Chapter 2 explores Fodor's (1998) compositionality principle, which states that meanings of a complex expression are determined by the meanings of its parts, in conjunction with some syntactic rules that govern their combination. Fodor holds that any theory of word meaning (such as the one expressed earlier in this chapter) needs to account for the compositional nature of language. Furthermore, he argues that the data motivating this principle are inconsistent with prototype theory. In response to this, it is argued that compositionality is a misguided principle, for it cannot account for referential meaning; where referential meaning is conceived as a specific form of meaning that involves a speaker picking out something in the external world, whether it is an object, action, or event. It is argued that it is only through grammar, the rule-based relations between words, that referential meaning is possible, and that individual words (or lexical items) are never referential on their own. In the final section of this chapter, this grammatical conception of referential meaning is linked to Vosse and Kempen's (2000) computational parsing theory of language (called their 'u-model'). This combined theory of grammar and parsing jointly explain important aspects of referential meaning and play a strong explanatory role in the Chapter 3.

Chapter 3 argues that certain forms of theorising – those that consciously or unconsciously assume aspects of the classical theory, described in Chapter 2 – lead to a specific type of ontological misclassification, termed 'noun phase (NP) reification', and to subtle, unintended changes in propositional meaning, termed 'clausal reification'. The former, NP reification, involves a specific type of a Rylean (1949/2009) category mistake in which an abstract noun is mistakenly taken to be of the concrete category. The latter, clausal reification, occurs when an abstract noun, embedded in an NP, occupies the subject position of a proposition, which is making an essentialist claim about an abstract noun/concept. It is argued

that clausal reification is especially problematic, because it makes a proposition, and theories based on these propositions, quasi-tautological; where 'quasi' here means partly/almost or similar to tautological.

These two forms of reification, NP and clausal, are framed as linguistic artefacts or linguistic biases, as unintentional products of the way human language functions. It is hypothesised that they occur for reasons relating to the construction of referential meaning. Building on the work in Chapter 2, it is argued that referential meaning is a grammatical phenomenon, and that different grammatical relations correlate with different referential strengths, therefore a referential-grammatical hierarchy can be described to capture these differences; developed in §3.3. When this hierarchy is considered in conjunction with Vosse and Kempen's u-model, they jointly predict that a word's/noun's lexical content (its meaning) is blindly integrated into already-established referential structure. Therefore, abstract nouns can enter referential configurations that are ill-suited for the fuzzy level of meaning stability they possess. This mismatch is the root cause of both NP and clausal reification. The details of these types of reification are too complex to describe here, but are fully explained in §3.5 and 3.6. On this point, it is concluded that the existence of these biases explains the diversity of seemingly plausible but contradictory positions documented on the self in Chapter 1.

The final part of this chapter positions these biases in the wider context of the cognitive bias literature. This literature holds that the human form of cognition is prone to make systematic errors. To date, language and grammar, as they are conceptualised in this thesis, have not been considered as an area vulnerable to cognitive biases. This discussion constitutes and attempt to begin this conversation.

In short, certain types of theorists (i.e. those that lean on aspects of the classical theory of concepts, and which make essentialist propositional claims on the nature of abstract concepts/nouns) are snared by a linguistically-orientated trap, a cognitive bias that subverts the meanings of their propositions, making them and their accompanying theories unintentionally quasi-tautological.

Chapter 4 – in response to the methodological critique presented in the previous chapters – advances a positive method for theorising on abstract nouns/concepts, demonstrated with reference to the self. In previous chapters and in this chapter, it is argued that abstract nouns are inherently unstable and need to be stabilised for them to be both useful and synchronised (in their use) across a linguistic community. By default, PLT (i.e. 'perceptuo-

linguistic triangulation', described in Chapter 2) is an inappropriate stabilising force, because abstract nouns are devoid of a clear third-point physical object in the world that would facilitate triangulation. However, it is argued that PLTb (i.e. 'perceptuo-linguistic triangulation of behaviour', also described in Chapter 2) may be an appropriate alternative, for abstract nouns can be linked to behaviours, which have an observable, shared existence in the perceptual domain and thus are susceptible to triangulation. Here, the use of the term 'behaviours' is used in a non-standard sense, referring to a broad range of actions that are produced by individuals, organisms, objects and/or systems, which all have an observable, shared existence in the human perceptual domain. This appeal to behaviour is advanced as an antidote to classical, essentialist theorising, which, as argued in Chapter 3, leads to quasi-tautological propositions and theories.

One of the major difficulties in utilising this PLTb approach, in building a PLTb profile for an abstract noun, is isolating the behaviours of interest (i.e. in this case, those indicative of self or selfhood) without resorting to the essentialist approach to concepts. Using PLTb, in conjunction with basic scientific methods and certain foundational assumptions in statistics – namely, the Gaussian distribution and central limit theorem – it is possible to circumvent this trap.

With this theory in place, the argument discusses which behaviours to integrate into the PLTb profile of self. In this first instance, the 1st person pronoun 'I', as one aspect of the grammatical person system, is considered as a possible behavioural indicator of self, since self-referring via this mechanism, is one of the defining characteristics of human communication, which, in its linguistic form, is exclusive to our species. However, after critical consideration, it is found that 'I', used in isolation, is largely inadequate. In its place, the production of fully socialised and statistically typical utterances are proposed as a core constituent behaviour in the profile of self; where 'utterances' are conceived as self-standing, syntactically independent, units of discourse, providing new information.

Next, this chapter addresses several possible objections to utilising this PLTb method. The last criticism discussed, on artificial intelligence, highlighted the need to integrate a behaviour associated with human comprehension into the PLTb profile of self, that is, in addition to language production (of utterances), discussed in the previous paragraph. It was proposed that this would be achieved via the triangulation of neuronal behaviour associated with the N400 and P600, electroencephalography (EEG) event related potentials (ERPs). Evidence suggests that these ERPs index semantic and syntactic comprehension, respectively. This discussion primes the empirical investigations in Chapter 5.

Chapter 5 presents two EEG experiments that explore the neuronal processes underpinning the language comprehension mechanism in humans, with a view to integrating this information – in conjunction with the previous literature – into the positive PLTb theory of self, outlined in the previous chapter. This chapter begins with an overview of the EEG technique. This is followed by a brief review of the EEG language comprehension literature, specifically that discussed with reference to the N400 and P600 ERPs, which, as stated above, are associated with the processing of meaning and syntax, respectively. These ERPs are the focus of the two EEG experiments presented in this thesis.

The first experiment is with neurotypical adults (N=26). The second experiment – which utilised the same experimental paradigm as the first – is with patients with a diagnosis of schizophrenia (n=11) and neurotypical adult controls (n=11). Patients with a diagnosis of schizophrenia were selected, because schizophrenia is often linked to a loss of self or attenuated selfhood. Therefore, it was of interest to explore to what extent that this is the case in view of this new PLTb approach to the self. Further information on the design and results of these experiments are not reported in this Preface, but overall, it will be concluded that aspects of neurophysiological evidence (including EEG ERPs), where the components can be observed and measured statistically, can be integrated into a PLTb profile. Here, it is recommended that the N400 and P600 ERPs be integrated in to the PLTb profile of self as triangulable measures of language comprehension, of meaning and syntactic processes. The final section of this chapter briefly discusses the consequences of this integration in the context of schizophrenia.

1 THE SELF IN PHILOSOPHY, PSYCHOLOGY AND NEUROSCIENCE

1.1 Chapter overview

This chapter retracks the initial observations that led to the meta-theoretical conclusions outlined in the Preface. Firstly, the reader is introduced to the concept of self, as it has unfolded in philosophy, psychology and cognitive neuroscience, presented in roughly-chronological order. In total, sixteen theories of self are outlined across these three disciplines, and at the end of each outline is a summarising proposition, which is intended to capture the essence of that particular position. These summarising propositions are intentionally articulated in such a way that the relevant author would be happy endorse these as representative of their position(s), or at least find them difficult to reject. The purpose of these summarising propositions will become evident towards the end of the chapter, after the initial review section. At this point, the various theories of self discussed in this chapter are examined from a birds-eye perspective, as a complete set of theories. From this viewpoint, it is determined that when these individual theories of self are considered as standalone, independent theories, they often prove to be internally consistent and persuasive, but when these theories are considered as a set of theories, significant contradictions arise that potentially undermine the analytical approach to concepts and their associated problems. This point is first developed in ordinary, argumentative prose and then in a more formal configuration, with logical terms. Ultimately, the meta-theoretical conclusion of this chapter motivates (in Chapters 2 and 3) a withdrawal from the debate of self, to consider some more foundational aspects of theorising, which might have led to this state of affairs.

The theories of self reviewed in this chapter have been selected because of their prominence in the field, as highly influential and discussed views, and as a representative sample of a range of ideas that have arisen in the history of western philosophy, in modern psychology and neuroscience. This review is by no means exhaustive, but of the theories not included, it is likely that these methodological criticisms will be applicable, for this approach to abstract concepts runs deep.

1.2 What is the self?

Loosely speaking, the branch of philosophy examining the self, sometimes synonymously referred to as ‘person(s)’, tries to outline the conditions of identity that make one subject of experience different from all others, at a particular time and across time. There is no single question associated with the self, but rather a wide range of connected questions, such as: ‘Who

am I?', 'What is it to be a self?' and 'What is it about a self that facilitates persistence through time?'

The first question: 'Who am I?' evokes answers in which an individual predicates certain properties onto themselves, such as: 'I am a man', 'I am writer', 'I am a philosopher', and so on… The second question: 'What is it to be a self?' gives rise to definitional/criterial answers, a list of things that the self is and is not. Questions and answers of this variety often consider the constitutive nature of a self unextended in time, and are often referred to as 'synchronic theories of self'. On the third question, regarding the persistence of self, we get further definitions, only this time geared towards aspects of the self that allow it to persist through time. For example, person X is the same person today as s/he was at some previous point in time, if-and-only-if certain necessary and sufficient conditions are satisfied (see §2.3 for more details on these conditions). Questions and answers on these aspects are often referred to as 'diachronic theories of self'' or as theories of 'personal identity'.

Below is an outline of the various philosophical answers to these questions on the self, starting with classical western theories, through to 21st century, modern philosophical incarnations. However, to stop at this point would constitute only a partial story of this concept, for in the late 19th century, interest in the self drifted outside of philosophy, firstly into psychology and then later into cognitive neuroscience. Consequently, the central answers from these disciplines are also outlined. Please note, however, that this particular review of the literature is more of a whistle-stop tour than a deep and considerate analysis of each individual position and response to it. There are strong reasons for adopting this minimalistic approach. The primary being that the argument under development here is attempting to highlight a very specific point about theoretical perspectives on the self, and about theoretical perspectives of abstract concepts, more generally.

1.3 The self in philosophy

Classical Greek (pre- and post-Socratic) theories of self are embedded and/or implied in theoretical views on other philosophical matters. Rarely are independent theories of self articulated for their own sake, and consequently, neither are they subjected to sustained critical thought (R. Martin & Barresi, 2006, p. 13). Plato's dialogues (429-347BCE/1997) were a slight exception to this, for in these we see discussion on content equivalent to that discussed in modern theories of self, only under the lexical guise of the soul/psyche. In Plato's *Phaedo*, for example, Socrates contends that the soul is immortal and immaterial, released from the body

at the moment of death. He suggests this transformation to be an act of purification, leading to unadulterated rationality and true knowledge of the forms; a conveniently timed set of propositions, given his impending hemlock encounter. This conception of the soul has textual support in other dialogues. For example, in the *Meno*, Socrates talks about the pre-existence of a soul prior to its incarnation into a body. Here, he uses this conception of the soul as a means of explaining a person's ability to acquire knowledge (e.g. knowledge is a form of remembering intellectual truths acquired before birth). This is not to say that Plato was completely consistent throughout his writing. There are contradictory conceptions of the soul in some of the later dialogues, namely, in *Phaedrus* and *Laws*. More important, though, than establishing Plato's exact view of the soul is acknowledging the influence he precipitated on this concept, as a precursor to this new branch of philosophy. Despite the minor inconsistencies, the following proposition captures the dominant view in the Platonic literature.

S_1 = the soul/self is immortal and immaterial

Like Plato, Aristotle (384-322BCE/2016) agreed that the soul was the essence of a human being, but largely opposed Plato's dualistic, two-worlds conception of the soul, as distinct from the body; but, as with all classical interpretations, this a debated interpretation (Barresi & Martin, 2013, p. 35). On Aristotle's account, the soul is linked to the body, being its form or essence, not a substance inside but distinct from the body. Aristotle's positive theory states that every organic thing has a soul and it is the soul that constitutes its vital principle or function. According to Aristotle, the structure of the soul is manifest in three different ways. First, the vegetative soul accounts for reproduction, and is present in plants, non-human animals and humans. Second, the sensitive soul accounts for perceptual processes, present in non-human animals and humans (i.e. plants are excluded). Finally, the rational soul (*nous*, in his terms), which accounts for reason and intellect, and is present in humans only, constituting our essential function. In further contrast to Plato, Aristotle showed little interest in individual souls, instead, arguing that the soul should be conceptualised in the context of human nature. Here, the focus shifts from the soul as a special part of the human organism, to souls (or selves) being only one particular manifestation of a more universal and immutable human nature.

S_2 = the soul/self is a form or essence of the being

In the period spanning the 1st to the 12th century, beginning from the Christian Neoplatonist philosophers, we see a narrowing of interest in to the survival aspect of the soul after bodily death, a focus not prevalent in their classical predecessors. On the Christian account, souls survive their bodily deaths, but they survive in some bodily way, such that resurrection after death requires the body and soul form an intimate unit. Augustine (354-430/2003) was especially influential in this period, and was one of the first thinkers to address the difficult relation between the soul and the body.

S_3 = the soul/self is immortal and immaterial, but intimately connected with the body

Moving forward to the 16th and 17th century, with its significant advancements in modern science, we see a new set of theories, which reconceptualise the self as part of a mechanistic universe. In this phase, the modern content of self is discussed with reference to the mind, and of this period, Descartes (1641/2008) was the first major thinker to start using the term as an alternative to the soul. On his view, the self is the mind, which is a thinking, non-extended thing; as opposed to the body, which he conceived to be a non-thinking extended thing. As one of the central thinkers in this mechanistic surge, Descartes' interest in the mind was to establish the extent to which it fits into the otherwise wholly mechanistic world and materialistic body, to see to what extent this thinking thing is independent of its ecological relations. For Descartes, the answer to this was that the neural organisation of the brain accounts for the various forms of perceptual processes that we experience, but that these function independently from the immaterial mind. On his account, only humans have a mind, non-human animals lack this immaterial selfhood, and as such, are just the product of predetermined, biological and mechanical operations. In the case of humans, the immaterial self and the perceptual processes experienced in the body are linked via a small endocrine gland in the brain known as the pineal gland. This relationship can be characterised as causally bidirectional, in that both the immaterial mind and the associated body can affect one another, but one particular immaterial mind cannot affect other material or immaterial things, besides the one associated body. For Descartes, this gives rise to the illusion that the mind and body are the same thing.

S_4 = the mind/self is immaterial, interacting with the body through an aspect of the brain

Towards the end of the 17th century, we begin to see discussion of the self of a more empirical and naturalised form, resembling the structure of present-day discussions. Locke (1689/1975) was the first philosopher to provide some content resembling the synchronic and diachronic distinction, discussed in §1.2. To reiterate, the former asks 'What does the self consist of?', the latter 'What allows the persistence of self over time?'. In answer to these questions, Locke argued for the view that the self, and its continuity over time, are a matter of biological continuity (continuous functional organisation) and psychological continuity (e.g. consciousness and memory), respectively. More specifically, Locke makes a distinction between a man and a person, linking biological continuity to the former and psychological continuity to the latter. According to Locke, a person is a rational thinking being and a person's identity over time consists in continuity and extension of consciousness and memory, forwards and backwards. Here, consciousness and memory are necessary and sufficient conditions for personal identity.

Post-Locke, this notion of an immaterial soul, whilst continuing to have a dominant existence in theology, largely dropped-off the radar in naturalised theorising, likely because of its lack of clear observable, triangulable, physical properties (more on this in Chapter 2: §2.2 & 2.3). Furthermore, Locke's perspective of the self, and the critical responses to it (Butler, 1975; Reid, 1850) pushed to the fore the idea that the self might be a fiction or illusion. For Locke, the self was an aspect of memory and consciousness, aspects of cognition that are intrinsically transient, in comparison to an immaterial soul, hence, a potential locus of vulnerability.

S_5 = the self is psychological continuity (e.g. consciousness and memory)

Hume (1739/1888) hit upon this transient nature, and consequently, rejected the notion of self persisting through time. Hume's method of inquiry begins from the assumption that all ideas are derived from impressions (or perception(s), in modern, simpler terms). Ideas of a persisting diachronic self, then, are derived from perception, but all perceptions are momentary, therefore, the idea of the persisting diachronic self can be considered a perceptual illusion, springing from the human habit of attributing unified existence to a collection of individual parts, in conjunction with our disposition to infer cause and effect. In different terms, when we introspect on our own experience of self, as persisting through time, all we identify is a

consecutive stream of inconceivably rapid and transient perceptions, much like watching a film or a play, with the coming-and-going of actors representing the transient perceptions. Here, perceptions are separate from one another and there is no unifying component of self that would signify permanence. What is left after Hume's critique, is a reduction of self to cognitive processes, (predominantly, perception) and when these perceptions are absent, as in sleep and in the death of the brain, the self is destroyed.

S_6 = the self is a bundle of perceptions

In the eastern tradition we see modern discussion on the self in the religion of Buddhism, discussion not dissimilar from Hume's (1739/1888) theoretical position. There are several interpretations of the Buddha's teachings, but typically they all deny the existence of a self as an ontologically existing thing and claim that it is the separate concept of person that is responsible for the illusion of self, as well as the experience of certain negative psychological behaviours.

According to Siderits (2013) interpretation of Buddhist reductionism, the self is conceived as a simple, basic ontological essence, whereas a person is the complex and total set of empirical/perceptual experiences that an individual experiences. Siderits refers to the latter as the 'psychophysical complex'. On his view, there is no self, because there are no enduring, ontologically simple entities, for all things are impermanent. Furthermore, it is the concept of person that gives rise to this illusion. A person, on this view, relates to the diachronic sense of self discussed above, and is presented as a conceptual and conventional fiction, which is useful to a point, but can be dangerous when taken too seriously, and is ultimately lacking in ontological permanence. Besides the similarities to Hume's work, there are similarities to Buddhist thought in Dennett's (1992) later-developed narrative theory of self, discussed below.

The illusion of personhood extended over time likely arises for very practical reasons. Specifically, it is beneficial to attribute permanence to a person, for it aides us in planning for a happier future, were our physical and psychological welfare is central to that future conception. As an example of this, Siderits asks us to consider the case of brushing our teeth (p. 302). Typically, children tend to dislike this task, but often they are instructed and motivated to stick with it, based on the assumption that if they brush their teeth regularly as children, then in the future, as adults, they will be spared the discomfort associated with tooth decay. In this

particular organisation of behaviour, there is a clear overall enhancement of future welfare, based on the conception of a person extended in time. However, on the Buddhist account, it is precisely this 'useful conceptualisation' where psychological problems in the individual may occur, especially when people take too seriously the idea that persons are ontologically real and extended over time. So, whilst there are beneficial aspects to this form of reasoning, as exemplified above, this form of reasoning may also lead to specific forms of psychological suffering, such as: alienation, frustration, and despair, through forcing us to confront the issue of our own mortality. In other words, the concept of a person is only useful to a certain point, and conceptualising about personhood beyond this point leads to specific forms of existential suffering. In response to this, the Buddhist argues that the ideal state is one in which we continue to behave like persons, with respect to the organisation of physical and psychological welfare, but to not take this idea of personhood too seriously, beyond these welfare considerations. In overcoming this ontological mistake, we can reap the benefits of thinking ourselves as persons, whilst simultaneously avoiding the aforementioned psychological suffering, by recognising persons as the ontological fiction they are.

S_7 = the self does not exist and a person is a conceptual fiction that is useful, up until a certain point

Kant's (1781/2007) remarks on the self are embedded in his more primary focus on understanding the operations of the mind and its limits with respect to knowledge. His style of reasoning and writing are notoriously intricate, therefore there is little consensus on how best to interpret his remarks on the matter. For certain, though, is the importance of Kant's distinction between the phenomenal and noumenal world. On Kant's view, the phenomenal world is the world of appearance, the world as we perceive it through our own particular cognitive phenotype and which constitutes our phenomenal experience. The noumenal world, in contrast, refers to the thing-in-itself or things-in-themselves, as constituents of reality considered in their 'real' form, independent of experience. Kant's synthetic *apriori/aposteriori* reasoning is only practicable at the phenomenal level. Reasoning about the noumenal realm would necessarily be independent of our experience of it; therefore, we are epistemologically ignorant of this realm.

One straightforward, non-metaphysical, interpretation of Kant's view, based on this distinction, is discussed by Barresi and Martin (2013, pp. 46-47). They interpret Kant as positing two selves, one relating to the phenomenal realm and one to the noumenal. The self of the phenomenal realm has spatial and temporal properties and can be experienced both subjectively and objectively. On this view, accounting for the objective aspect of the self is no different than accounting for any other object that exists and has spatio-temporal properties. The self of the noumenal realm, in contrast, lacks or transcends these properties and therefore, cannot be known, and is thus considered ineffable and indescribable, yet all the while is instrumental in the structuring of the phenomenal world. For a more detailed interpretation of Kant's philosophy of self, see Melnick (2009) and Brook (2016).

S_8 = The self is not an object of experience, but is a transcendental condition for it.

Gergen (2013) diverges from the traditional Western conceptions of self as fundamentally individualistic, acknowledging a primacy of the social over the individual, considering the self to be a construct of social relations. On this constructivist account, Gergen argues that knowledge, a primary aspect of the mental realm, is not part of the individual mind, but is embedded in the communicative relationships between humans. Therefore, knowledge and other aspects of the mental realm are all subject to, and governed by, specific linguistic conventions and/or language games (Kuhn, 1970; Wittgenstein, 1953/2009). On this communal view of knowledge – and mental contents, more generally –no one arrangement of words function as a better depiction of reality than any other, they are all meaningful within their own respective traditions. Here, a specific conception of language is advocated where the meaning of concepts is a derivative of the particular use of a language embedded within social relations. From this, it follows that the concept of self is also a derivative of language use, which is why the self is conceived as a social construct.

S_9 = the self is a social construct, residing in the communicative relation between individuals

As alluded to above, Dennett (1992) develops a narrative theory of self, in which the self is conceived of as an abstract thing or as a convenient psychological construct, as opposed to

something tangible or ontologically real that can be found in the world. To demonstrate this point, Dennett compares the self to the centre of gravity. The centre of gravity is an abstract theoretical concept hypothesised in physics. It is not a physical item in the world, it has no spatio-temporal properties, but it does have a well-delineated, mathematically defined, role in physics and is useful in explaining the behaviour of objects, especially at the macro level. As an example of this, Dennett initially discusses the centre of gravity with reference to a chair. On this, it can be said that humans are highly attuned to the fact that a chair has a centre of gravity, because if we were to gradually tip a chair backwards, we would be able to discern the point of no return, were the chair would fall.

Following on from the above example, Dennett then discusses the centre of gravity of a complex machine, an engine of some variety in which the centre of gravity moves in some systematic and uniform way as the engine runs, say, as the pistons oscillate and rotate. He asks us to imagine that we have some sufficiently clever technology that could track and plot the movement of the centre of gravity in this complex machine, and he asks us to suppose further, that the movement of the centre of gravity perfectly correlates with some physical aspect of the engine, namely, a particular iron atom. It is Dennett's contention that even if we discovered that the centre of gravity perfectly correlated with a particular iron atom of the engine, we could not say that the iron atom and the centre of gravity are identical. To make such a statement would constitute a category mistake, because the centre of gravity is not in the same ontological category as an iron atom; the latter has spatio-temporal properties, whereas the former does not (see Ryle, 1949/2009 and §3.2, for more details on category mistakes).

Dennett considers the self to be in the same category as this centre of gravity concept. Therefore, when theories of the self ascribe spatio-temporal properties to the self they are making category mistakes, equivalent to those outlined above. This ontological bad habit, according to Dennett, arises in the human attempt to explain complex behaviour. By analogy, the physicist in calculating how an object moves, posits this theoretical centre of gravity to assist in the calculation of its behaviour in the world. In like manner, self theorists, in observing a variety of biological things moving around in the world (in particular, humans and non-human animals), postulate the existence of selves and ascribe a narrative to them, in order to interpret and explain their complex behaviour. Dennett constructs several hypothetical examples to demonstrate that the existence and construction of human narratives does not require an ontologically existing, physical, self for these to arise. As such, all that is left is the narrative.

S_{10} = the self is a narrative fiction

1.4 The self in psychology

Psychology was considered a branch of philosophy for many centuries, as is implicitly evidenced in the discussions above, with their overlapping interest in mental states and behaviour. This continued until the late 1870's, at which point, Wundt opened the first laboratory exclusively dedicated to psychological research. This action was crucial in enabling psychology to evolve from being a predominantly philosophical enterprise into an independent scientific and clinically-orientated discipline (Benjamin, 2000). In the same period, James (1890/2007) integrated his own naturalising impulse into psychology, incorporating philosophical, scientific and clinical observations into his theoretical perspective of self (Leary, 1990).

James defined the self in two different ways, the self as knower (the transcendental subject: I, or the pure ego) and the self as known (the empirical object: Me). The latter, the self as known, is broken down further into the Me viewed as material, the Me viewed as social and the Me viewed as spiritual, where the self (as known) is the sum-total of all these aspects. The material aspect of self is not a standard materialistic conception of self that one would expect, as in philosophical materialism or physicalism. Instead, for James, the material aspect of self consists of things that belong to the individual person, but this belonging is an emotional relation, as you might find between a person and one's body, clothes, family, friends, property etc. The social aspect of self relates to the innate propensity for a person to have social relations with other individuals e.g. felt relations. Finally, the spiritual aspect of self was developed to acknowledge the fact that we can think of ourselves as thinkers. This is not to be confused with the ontologically pure ego which he barely speculates on (Leary, 1990, p. 111), likely because he thought it to be unnecessary for the science of psychology to proceed.

S_{11} = the self has four aspects; material, social, spiritual and pure ego

Approximately 100 years later, Neisser (1988) offered another divided theory of the self. He argued for a five-way division, which emerges at different stages during typical human development. Each of these aspects are distinct from one-another, to the extent that they can

be considered as different selves in the context of developmental and cognitive profile, as such, these different aspects are subject to different pathologies.

In capsule form, these selves include the ecological self, which appears during early infancy, perceived with respect to personal spatial awareness in the immediate physical environment. This involves unreflective awareness that we possess a bounded, articulated and controllable body that processes kinetic and optical information, and furthermore, that this body interacts with an objectively existing environment. Though present in early infancy, this form of self fully develops with increasing age and skill.

The Interpersonal self, also appearing in infancy, involves the signalling of emotional rapport and communication between humans. This intersubjectivity is unreflective and immediate and occurs in social interactions, initially between mother and child, but later between two or more children/adults. These intersubjective communications are genetically endowed, but require a peppering of behavioural learning. These forms of communication are manifest in facial expressions, gesturing and vocalisations.

The extended self, relates to episodic memory of personal experiences, and to the memory of routines that we engage in. With respect to episodic memory of personal experiences, this involves the ability to recollect unique and past events; in the first person (e.g. a recent holiday), whereas memory for routines relates to more general event representations (e.g. I am an individual who occasionally goes on holiday). The extended self is the cumulative total of these two different kinds of memory.

The private self, develops when a child first discovers that some of their conscious experiences are not shared by others. This include inner aspects of perception, such as dreaming, thinking, pain, hunger etc. This form a self plays an important role in self-knowledge.

Finally, we have the conceptual self, which develops when we are able to acquire concepts and attribute them to ourselves e.g. I am a man, I am a writer, I have a liver, I have certain social rights, and so on… As we develop linguistic competency, each of us is able to develop a conceptual schematic of ourselves.

S_{12} = the self has five aspects; ecological, interpersonal, extended, private and conceptual

Gallagher (2000), states that recent approaches in psychology and philosophy can be divided into two categories, focusing on two important aspects of self that together may lead to a fully comprehensive theory, these are the narrative self and the minimal self. Gallagher's reference to the narrative self refers (approximately) to the aforementioned work by Dennett (1992), about past and future stories that individuals construct about ourselves. Whereas, his reference to the minimal self relates to the various attempts to establish the minimal requirements of selfhood. Gallagher argues that to establish these, all unessential features of the self should be stripped away, and whatever is leftover constitute the bare necessities of a self, i.e. the minimal self.

For Gallagher, the minimal self is a phenomenological consciousness of oneself as the immediate subject of experience, unextended in time (2000, p. 15). This particular conception of the minimal self is conceived as ecologically embodied, is dependent on specific brain processes and possesses two key characteristics: a sense of ownership and a sense of agency; developed in the context of motor action and perception, as detailed by Frith (1992). A sense of ownership and sense of agency are barely distinguishable in the normal experience of action and perception, in that we typically experience both senses at the same time. For example, when I reach for a glass of water, I know that it is me orchestrating this voluntary agential action (agency) and that it is my body involved in the reaching for the glass (ownership). In this instance, I acknowledge both the agency and the ownership of the movement. However, it is possible to have ownership without agency, in that I may acknowledge ownership of a movement, whilst not experiencing a sense of agency. For example, when a physician uses a reflex hammer during a medical examination, and the hit limb twitches. For Gallagher, it is these two aspects that constitute the minimal self (i.e. an appropriate sense of agency and ownership) and it is these facets – amongst other things – that are said to malfunction in schizophrenia (Frith, 1992).

S_{13} = the self is a sense of agency and a sense of ownership functioning correctly, in conjunction with a narrative

Churchland (1981) developed the position referred to as 'Eliminative Materialism'. This is the view that our common-sense conceptions of psychological phenomena (i.e. our folk psychology concepts) are radically false. Moreover, it is Churchland's contention that these

folk psychological theories, with their associated ontological commitments, will eventually be eliminated and replaced with a neuroscientific paradigm, informed by modern neuroscientific evidence. It is expected that this neuroscientific theory will have more explanatory power than the common-sense psychology it displaces, and it will be more substantially integrated with the other physical sciences.

Applying this eliminative theory to the self, would most likely lead to the view that the concept of self is a presupposition in folk psychology. Therefore, the failures to solve the problems associated with the self are indicative of a larger-scale problem with the underlying theory, namely, folk psychology. Given that folk psychological theories have their roots in earlier philosophical work, it follows by inference that certain aspects of philosophy will also fall by the wayside when neuroscience reaches a certain level of advancement.

S_{14} = the self is a folk psychological concept, which will ultimately be displaced and explained through neuroscience

1.5 The self in cognitive neuroscience

Cognitive neuroscience is a recently founded scientific discipline, which attempts to establish the neural mechanisms that underpin different aspects of cognition. Here, 'neural mechanisms' refers to the behaviour of neurons and neural networks. Neurons are the major cell component of the central nervous system (i.e. the spinal cord and brain) whereas neural networks refer to the passing of information between neurons, through electrical and chemical signalling. The modifier term 'cognitive', in 'cognitive neuroscience', refers to the 'variety of higher mental processes such as thinking, perceiving, imagining, speaking, acting and planning' (Ward, 2015, p. 2); the topics of interest in mainstream cognitive psychology. Cognitive neuroscience, then, is the combined discipline of neuroscience and cognitive psychology.

Cognitive neuroscience only established itself as a stable scientific discipline in the last 40 years, from the 1980s onwards (Raichle, 1987/2011). This stability, and subsequent popularity, of this discipline was driven by advances in brain imaging technology, such as: positron emission tomography (PET), functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and magnetoencephalography (MEG), amongst others. On this subject, note that specific details of the EEG methodology will be explained in §5.2, as a precursor to the experimental EEG research of this thesis, in §5.6 and 5.7.

Most, but not all, imaging studies rely on the cognitive subtraction method as a means of mapping different forms of cognition to neuroanatomy. Cognitive subtraction is an aspect of experimental design, which involves a particular type of contrast between two conditions. What is special about this design is that these two conditions only differ in exactly one cognitive respect (Harrison & Pantelis, 2010). As a simplified hypothetical example of this, imagine that Condition1, of some basic two-condition-imaging experiment, engages the cognitive processes: A and B, therefore, [Condition1 = A & B]. Whereas, Condition2 engages the cognitive processes: A, B and C, therefore, [Condition2 = A, B & C]. It is this one difference (the 'C' of Condition2) that is the cognitive feature of interest for the experimenter, and is therefore, the independent variable (IV) of interest in the experimental design. To tap into this IV, the neuronal activation (i.e. the dependent variable) elicited by Condition1 is subtracted from the activation elicited by Condition2, leaving only the activation associated with C of Condition 2. This cognitive subtraction technique presupposes 'pure insertion' as a viable methodological possibility in the mechanics of neurological activity. This presupposition refers to the belief that a specific cognitive process (e.g. 'C', of the previous example) can be inserted into a task, without it affecting or interacting with the already present processes (e.g. 'A & B'). Whether this is possible is still a matter of debate; see Friston and Price (2011) for overview.

With respect to the self in cognitive neuroscience, what we see are experimental tasks that contrast self-related processing with non-self-related processing, with the aim of identifying the neural correlates of self across different contexts and modalities. Typically, the cognitive subtraction method is utilised with the self-related element functioning as the IV.

Example paradigms that contrast the self and non-self include: perceiving one's own face compared to other faces or warped faces (Kircher et al., 2001); the processing of one's own name, in comparison to other proper names (Perrin et al., 2005); the processing of personal pronouns, compared to non-personal pronouns (Walla, Greiner, Duregger, Deecke, & Thurner, 2007); the judgment of personal traits and characteristics (Craik et al., 1999; Ochsner et al., 2005) and in the attribution of self-initiated action (Farrer et al., 2003), to mention just a few. The discussion below will help guide the reader to a more comprehensive list of research on the topic.

One particularly influential model of the self that has arisen out of this influx of research, states that the cortical midline structures (CMS) of the brain are fundamental to the constitution of self. At the forefront of this modular conception of the self is Northoff and colleagues: (Northoff, 2011, 2013; Northoff & Bermpohl, 2004; Northoff et al., 2006).

Northoff et al argue that the self can be identified with (or equated to) self-referential processing in the brain; where self-referential processing refers to the neurological processing of stimuli that are 'strongly related to one's own person' (Northoff, 2005, p. 211); with 'strongly related' referring to external stimuli that have a strong association with the individual person. On this view, self-referential processing in the CMS represents the core of what the self is.

To demonstrate this point empirically, Northoff et al. (2006) performed a meta-analysis of 27 imaging experiments on self-related tasks, published between the years 2000 and 2004. They found that sub regions within the CMS correlate with seven cognitive domains involved in self-related tasks, in the self-condition. In particular, they demonstrated that the self-referential processing in the verbal domain occurs in the ventromedial prefrontal cortex (Johnson et al., 2002; Kelley et al., 2002); in the spatial domain, in medial parietal cortex and posterior cingulate cortex (Marshall & Fink, 2001; Vogeley & Fink, 2003); in the memory domain, in dorsomedial prefrontal cortex (Macrae, Moran, Heatherton, Banfield, & Kelley, 2004); in the emotional domain, in ventromedial prefrontal cortex and pre- and subgenual anterior cingulate cortex (Phan, Wager, Taylor, & Liberzon, 2004); in the facial domain, in medial anterior cingulate cortex (Kircher et al., 2001); in the social domain, in anterior and posterior CMS; anterior cingulate cortex and medial parietal cortex (Frith, 2002; Kampe, Frith, & Frith, 2003; Vogeley et al., 2001); and in the agency and ownership domain, in anterior insula (Farrer et al., 2003; Farrer & Frith, 2002).

In addition to these findings, Northoff et al. suggest that the cortical midline structures are connected to subcortical midline structures, to form a cortical-subcortical midline system of self. On this view, the subcortical system gives rise to the bodily representation of the individual, that is, the schematic we have of our own bodies marked in 'viscero-somatic motor coordinates' (Northoff et al., 2006, p. 449).

S_{15} = the self is an individual's relation to external stimuli in the world, mediated through self-referential processing. This processing is a functionally independent module in the cortical midline structures of the brain

Another influential neuroscientific theory of self was developed by Metzinger (2003, 2008). He argues that there are no such things as selves, in the substantial, ontological sense,

as things possessing self-sufficient existence in the world. Nevertheless, he does argue that there are such things as 'phenomenal self-models' (PSMs). On Metzinger's view, PSMs are continuously updating, dynamic self-representations that emerge from complex, information-processing mechanisms and representational processes, rapidly unfolding in the central nervous system (CNS) (2008, p. 216). In other terms, PSMs can be considered as schematic representations of the self, which generate the first-person perspective we consciously experience, as a product of genetically hard-wired neuronal activity occurring in the CNS.

These self-models are present in systems where there are intentional interactions with the physical environment, with the human-specific, conscious form of PSM likely arising as a product of evolutionary pressures in the cognitive arms race. According to Metzinger, it is the characteristics of these human specific PSMs that give rise to the illusion of ontologically-existing selves. Specifically, Metzinger argues that self-models are transparent, in that the information processed in the CNS does not reach the level of phenomenal content; instead, we look through it. To put it another way, because we do not have access to the neuronal processes that are occurring in the CNS, which give rise to our conscious phenomenal experience, we tend to posit the existence of a mysterious self entity to account for this form of experience. We confuse the content of PSM's with the existence of some Cartesian, immaterial self (described in §1.3).

To support this position, Metzinger utilises compelling arguments and evidence from philosophy and neuroscience, respectively. From philosophy (2003, Chapter 8) he develops arguments and metaphors based on Plato's cave and on flight simulators, to explain the nature of the mind and self. From the field of neuroscience (2003, Chapter 7; 2008) he leans on neurophemonological studies, such as that found in the rubber-hand illusion (Botvinick & Cohen, 1998), in phantom limb syndrome (Ramachandran & Rogers-Ramachandran, 1996), in out-of-body experiences (Metzinger, 2005), in evolutionary and artificial robotics (Bongard, Zykov, & Lipson, 2006; Maravita & Iriki, 2004) and in virtual reality research (Metzinger, 2008, pp. 233-234).

S_{16} = the self does not exist, it is an illusion brought about as a consequence of certain processes in the brain. However, a self schematic does exist.

1.6 The problems with multiple theories of self

When reading and considering the truth and falsity of the theories outlined above [S_1 to S_{16}] and others not mentioned, it is often the case that many of these theories make relatively good sense when read in isolation: they read as persuasive and their conclusions appear to follow from true premises. However, it cannot be the case that all of these positions are true, because when we consider these individual theories as a set of theories, and make comparisons across the set, it quickly becomes apparent that many of these theories are contradictory to one another. For example, depending on which, if any, of the above perspectives you accept, the self may be immaterial or material, permanent or transient, knowable or unknowable, singular or multiple, individual or social, phenomenal or neuronal and so on. To my knowledge, Olson (1998) was the first to hit upon this lack of agreement in the philosophical literature of self, and when we add the psychological and neuroscientific theories into the set, the potential for contradiction grows exponentially; although, a subset of theories are arguably compatible with one another.

As one example of contradiction, let us consider the above mentioned 'permanent-transient' disjunction with reference to Descartes' and Hume's respective theories of self. Descartes' position is that the self (as in 'mind') exists and is an immaterial, permanent, mental thing, distinct from the body. Whereas for Hume, there is no self in this mental sense. All there is, is a complex bundle of transient perceptions that give rise to this Cartesian illusion. Here, it is relatively clear that these two positions are at odds with one another, they both cannot be true. Yet, both theories, considered individually, are well argued for and make relatively good sense.

As mentioned above, besides the several contradictory positions, there is also a subset of theories that are compatible with one another. For example, Aristotle's and Kant's positions are potentially compatible with Metzinger's, only in different respects. Aristotle argues that the self is the essential form of a thing, whereas Kant argues that the self is not an object of experience, but a condition for it. For Metzinger there are no such things as selves, stated very much in the Kantian spirit, but there are such things as phenomenal self-models, which are similar to Aristotle's conception of essential forms. However, despite this overlap of agreement, there are also elements within these three theories that are contradictory.

From within the S_1 to S_{16} list above, there are 120 possible comparisons of paired theories, and given the disparate nature of the theories outlined above, it would be fair to

assume that a large proportion of these theories are either partly or wholly contradictory with one another. If we accept this proposition – that there are contradictions between these theories – then it must follow that not all of these theories can be correct, because that would violate the law of non-contradiction:

For example, imagine for argument's sake, that S_1's content negates the propositions in S_2, S_3, S_4…

$$S_1 \Rightarrow \{\neg S_2, \neg S_3, \neg S_4 \ldots\}$$

Therefore, S_1 is not equivalent to S_2, S_3, S_4…

$$S_1 \neq \{S_2, S_3, S_4 \ldots\}$$

And, for argument's sake, we assume that S_1 is true:

$$S_1 = \text{true}$$

Then S_2, S_3, S_4… cannot be true:

$$S_2, S_3, S_4 \ldots \neq \text{true}$$

This formalisation demonstrates the largely obvious point that contradictory theories cannot all be true. In particular, it highlights the fact that there are a plurality of plausible but contradictory theories that exist on the self, that is, assuming one accepts the premise that the multiple theories of self are individually plausible but largely contradictory with one another. Hereafter, for efficiency's sake, this latter point shall be referred to as 'PCT': **P**lurality of plausible but **C**ontradictory **T**heories that exist on the self.

One could make several replies against this PCT point. A likely response – with respect to the question of: 'What is the self?' – is that whilst the theorist agrees with the above

formulation, s/he also believes that one particular theory of self is correct. The problem with adopting a 'one theory solution' is that the individual would have to demonstrate how and why this one theory should be considered correct, above all others, when these 'others' are themselves argumentatively persuasive.

One way to support this one theory perspective would be to produce a set of theoretical arguments in support of the alleged correct position, and to produce theoretical arguments against contradictory positions. This has been the principal method in analytical philosophy since Plato. However, there are several problems associated with this approach that will be fully detailed in Chapters 2 and 3. For the purposes of this chapter, it is sufficient to point out that this approach has been largely unsuccessful for more than two millennia, in the sense that there is little to no agreement/consensus on the nature of abstract concepts and their associated problems, especially if we compare this progress and consensus with that made in certain branches of the sciences (e.g. medicine, physics, engineering, etc.). In Chapters 2 and 3, it will be argued that this lack of progress is indicative of a flaw in methodological approach.

The one theory solution is not the only attempt to account for the PCT conundrum. Gallagher (2013) proposes a different method, which he calls the pattern theory solution. Gallagher recommends that we accept the above-mentioned plurality of incompatible theories as a completely unproblematic fact. The main idea behind pattern theory is that the concept of self, rather than possessing some essential property captured by one theory, instead should be conceived as possessing a complex and jointly sufficient pattern of hierarchically structured contributories (p. 3), were 'contributories' refers to the various individual theories of self or specific aspects of these individual theories (note that in the context of Gallagher's pattern theory, the terms 'theories and 'aspects' will be used interchangeably). On Gallagher's view, these aspects are hierarchically structured, because they can have different weightings in the pattern theory of self, with some aspects more essential and heavily weighted than others. Principally, what this means is that none of the individual theories of self, such as those described above (S_1 to S_{16}), are wholly and ontologically/epistemologically correct in-and-of-themselves. Instead, in order to account for the self, it is necessary to consider it as a cluster concept, where a collection of these theories, configured in a certain hierarchical pattern, are considered individually necessary and jointly sufficient to account for what the self is. On this view, 'selves operate as complex systems that emerge from dynamic interactions' of constituent aspects' (Gallagher, 2013, p. 3).

In order to fully explicate this pattern theory of self, Gallagher reproduces the structure from an already-established pattern-model on emotion (Izard, 1972), applying the foundational principles of this model to the concept of self. This begins with Gallagher outlining a list of possible aspects that might factor into the cluster concept of self. This list included: *minimal embodied aspects*, relating to theories that consider the self from an embodied biological perspective – *minimal experiential aspects*, relating to theories that consider the self from a phenomenological, pre-reflective, first person, agential perspective – *affective aspects*, relating to theories that put emphasis on the role of emotion in the constitution of a self – *intersubjective aspects*, relating to theories that put emphasis on intersubjective attunement between humans in the constitution of a self – *psychological aspects*, relating to theories that stress the importance of memory and psychological continuity of personality in the constitution of self – *narrative aspects*, which emphasise the role of narrative in the construction of self – *extended aspects*, were the self is said to extend beyond the traditional conceived limits of the mind/body to physical objects that the person owns or is strongly associated with – and finally, *situated aspects*, which emphasise the role of environmental structures in the construction of self and selves (e.g. power structures, linguistic communities, etc.).

Here we have a list of eight different theoretical avenues on the self, each of which champions a specific aspect of this concept. It is Gallagher's contention that some pattern of these individual theories satisfies the conditions for what the self is. To explain this idea in a different way, imagine the hypothetical process of building/baking a self from scratch. On Gallagher's perspective, some pattern of the above-mentioned individual aspects, and potentially others not mentioned, would be necessary and jointly sufficient ingredients in the construction of a self. It is not necessary that all of these aspects be present, just some permutation that is sufficient to pass the threshold of what it means to be a self.

The motivation for this pattern theory is twofold: Firstly, it claims to account for the PCT problem. On this point, Gallagher holds that the self is some pattern of theories, not one individual theory, so the fact that there are many plausible theories fits perfectly with his model, because a sufficient pattern requires several plausible theories as constitutive parts. The further issue, that some of these plausible theories contradict one another can also be circumvented on this pattern account. This is achieved by simply stating that particular incompatibilities will not figure in the same pattern. Alternatively, one might argue that because different aspects can possess different weightings in a pattern – as part of Gallagher's hierarchical structure

presupposition – it is conceivable that incompatible aspects could be included in the same pattern, made permissible through differences in aspect weighting.

The second motivation behind the development of this position is to account for the neuroscientific evidence on the self; specifically that produced by Northoff et al (2004; 2006), discussed above (§1.5). To reiterate, Northoff's research states that there is a commonality uniting the different aspects of the self, namely, the processing of self-referential stimuli, across a diversity of contexts and modalities. Northoff argues that these various forms of self-referential processing correlate with specific neuronal activity in the cortical midline structures (CMS) of the brain, with different cognitive aspects of self-processing located in different areas of the midline structure. Here, the CMS function as the overall module for the self, but also, areas within the CMS are further specialised for specific forms of self-processing. In this respect, the empirical data fits Gallagher's theoretical formulation, because both suggest that the self is made up of various aspects that function together as a pattern. On this perspective, when these aspects do not function optimally (i.e. when a specific area or areas of the CMS is/are damaged) it is predicted that the individual will experience a diminished form of selfhood, and there is empirical evidence, produced by Northoff and others, to suggest that this is the case.

Unfortunately, there are several problems with this pattern theory solution, some of which undermine the proposed advantages outlined above. Responding to these advantages in turn, the first problem with this pattern model is that it does not actually solve the PCT problem, it only delays its onset. To recap, this problem refers to the fact that there is a plurality of plausible but contradictory positions that exist on the concept of self. Accepting the pattern theory solution, requires that we hold on to the view discussed in previous paragraphs: that lots of plausible theories are required for a pattern and that incompatibilities can be circumvented. At a glance, the pattern theory does appear to address the issues related to the plurality problem, but unfortunately, the very same issue immediately resurfaces when we engage critically with the consequences of holding to a pattern solution. For example, when we try to establish which of the specific patterns is ontologically correct, when we try to find epistemological justification for a pattern's inclusion/exclusion criteria (e.g. which aspects to include and how many), and when we try to establish epistemological justification for decisions of hierarchy, etc. In all of these cases, we are bereft of valid solutions. Gallagher himself does not attempt to answer any of these questions, he merely acknowledges it as an outstanding issue for future research (p. 4).

A second issue with this theory relates to its attempt to account for the empirical data on the matter of self, specifically work produced by Northoff et al (2004; 2006). On this point, it is worth mentioning that a number of recent review articles have called into question this specific line of research, critically outlining both conceptual and methodological flaws in the idea of a modular system in the brain exclusively allocated to the self (e.g. the CMS). For example, Gillihan and Farah (2005), like Northoff, also performed a meta-analysis of the self literature over a broad range of different aspects that contrast the self and non-self, such as: face recognition, body recognition, agency, personal trait judgments, autobiographical memory, and first person perspective. Based on this review, Gillihan and Farah determined that there is no clear pattern of neurological localisation in these self-orientated paradigms. Furthermore, from within these paradigms, Gillihan and Farah highlight several confounds in experimental stimuli and paradigm design. Consequently, they conclude that theories that attribute a neurological, unitary modular system of the self in the brain are doing so falsely, or at least prematurely, based on the diversity of non-modular activations present in the literature.

In another comprehensive meta-analysis, Legrand and Ruby (2009) demonstrated that self-relatedness activity involves a wide cerebral network, which they refer to as the 'Evaluation Network' or 'E-Network', for short. The review they performed implied that the regions underpinning the E-Network, across the studies of self, have a strong overlapping resemblance to the neurological correlates active during theory of mind (ToM), were ToM refers to the ability of an individual to attribute and understand mental states of others. Stated differently, Legrand and Ruby show that the main areas of the brain involved in the processing of self-related information are precisely the regions involved in ToM, so there is no preference for the self in these areas; no neurological dissociation. In its place, Legrand and Ruby demonstrate that activity in this E-Network can be explained by the involvement of other cognitive processes that are common to all tasks that recruit inferential and memory processing.

In short, the pattern theory solution to the PCT problem is unsatisfactory. That is, a pattern theory solution does not help us to differentiate between the multiply plausible, but contradictory theories of self that exist. Furthermore, it is not clear that the empirical data invoked to support this theoretical position is reliable, in that it is still very much an open question as to whether there are structures in the brain specifically dedicated to the self. Finally, related to this point, it is not at all clear that the term 'self', in these various neuroscientific paradigms, is being used in a uniform and commensurable manner, more on this in Chapters 2 and 3.

In light of these various failures to the PCT problem, the argumentative focus hereafter takes a step back from the frontline debate of what the self is, to explore more foundational reasons for why these multiply persuasive, but often contradictory theories arise in the first place; in the self, but also in other abstract concepts. This Cartesian-regress manoeuvre, to more foundational questions, begins with an examination of how words are acquired, how their meanings are structured, and how they are used to refer to things, and it is here where the criticism of the one theory solution is fully explicated (Chapter 2). Specifically, it is argued that this situation has arisen as a by-product of inappropriately applying the classical/analytical methods of analysis to the examination of abstract concepts. In participating in this approach, it is argued that a theorist unconsciously and unintentionally slips into a world of reification (Chapter 3), likely due to the existence of certain cognitive biases (§3.7). Consequently, theories allowing itself these slips (i.e. those consciously or unconsciously utilising analytical methods to explain the nature of abstract concepts, such as in S_1 to S_{16}, plus many others) engages in reification, and inadvertently develops a theory which is symptomatically tautological (§3.6). In simpler terms, the multiple and seemingly-contradictory theories of self make sense, because they are talking about (or referring to) different things, in which they are specifying their own terms.

1.7 Chapter conclusion

To conclude, the first aim of this chapter was to re-track the prominent philosophical, psychological and neuroscientific theories of self. This review aspect unfolded in roughly chronological order, starting in philosophy with: Plato, Aristotle, Augustine, Descartes, Locke, Hume, Siderits, Kant, Gergen and Dennett, then moving onto psychology with: James, Neisser, Gallagher and Churchland and finally, to the consideration of two dominant theories in cognitive neuroscience with: Northoff and Metzinger (see Table 2 in §3.6 for summary of the positions). This minimalistic review was performed primarily as a means of highlighting a specific methodological consequence of adopting a certain approach to abstract concepts, namely, that it leads to the development of multiply plausible but contradictory theories of self. Two possible solutions to this situation were then examined. First, the ‘one theory solution’, which states that of the multiply plausible but largely contradictory theories, only one is correct, so there is no methodological problem. This position is only briefly considered, specifically, highlighting its poor record of accomplishment in achieving consensus, despite its long dominance as an approach. This superficial critique is deepened in Chapter 2. The second

position considered was the ‘pattern theory solution’, which states that the plurality of theories is unproblematic, because a pattern of theories is required to explain the nature of the self. Unfortunately, this approach only delayed the onset of the problems outlined in §1.6. It begs several questions on how to establish the correct pattern of theories, how to distinguish between competing patterns, how the hierarchy of individual theories is established and how all of these decisions are epistemologically justified? As discussed in the previous paragraph, these factors drive subsequent discussion, in Chapters 2 and 3, into more foundational and metatheoretical areas. This will begin with a consideration on the nature of how words – in particular, nouns – work.

2 FROM WORD MEANING TO GRAMMATICAL REFERENCE

2.1 Chapter overview

In the previous chapter, literature on the self was reviewed from a philosophical, psychological and neuroscientific perspective. It was concluded that there is a methodological problem with the way in which the self, as an abstract concept, is being approached. This motivated a step-back from the frontline (first-order) debate in this chapter, to consider more foundational and metatheoretical areas relating to words (in particular, nouns) and how they represent concrete and abstract concepts.

This chapter begins with an examination of how the meanings associated with concrete and abstract nouns arise in humans. It is argued that the meanings associated with these two types of nouns likely arise in different ways and are subject to different cognitive processes. After establishing this point, the focus then moves on to how the information associated with these two types of nouns is internally structured, with both classical and statistical theories considered. It is concluded that the statistical conceptions of this information structure – such as that proposed in prototype theory – best explain the currently available empirical data on the matter, as well as satisfying important theoretical necessities associated with word/noun meaning, namely, the criteria of stabilisation and referential range. Moving beyond individual nouns, the argumentative thread engages with another theoretical necessity associated with word meaning, namely, Fodor's compositionality constraint, of how word meanings combine with one another to make more complex meaning. In response to this it is argued that this conception of compositionality fails to explain referential and propositional meaning, two fundamental aspects of human cognition and language. In the wake of this criticism, it is argued that it is only through grammar that referential and propositional meaning become possible. In the final section of this chapter, this grammatical conception of referential meaning is linked to Vosse and Kempen's (2000) computational parsing theory of language. This combined theory of grammar and parsing, jointly explain important aspects of referential and propositional meaning and will play an important explanatory role in subsequent chapters.

2.2 Meaning in concrete and abstract nouns

Only a subset of words name everyday physical objects. In linguistics, these words are referred to as concrete nouns and include such things as: 'trousers', 'motorbikes', 'bananas', 'plants', 'cats' and so on. These words are typically acquired ostensively, through perceptuo-linguistic triangulation (PLT), where there is a three-fold, cross-modal, perceptual interaction between two or more individuals and an object in the world (Bloom, 2000; 2001 alludes to an interaction

of this form, but is not of central focus to his line of research). For example, imagine a teacher drawing the attention of a child to an object that the child had never previously experienced, and then, during mutual perceptual attention, the teacher utters the objects' name, in order that the child may associatively learn the word-object relation. Through repeated instances of these PLT interactions, advanced conceptual information for these objects develop and are stored in the brain as memory, associated with the relevant word (Davis & Gaskell, 2009; Huth, de Heer, Griffiths, Theunissen, & Gallant, 2016). This form of information can be considered 'advanced', in that it encodes information beyond that found in pre-linguistic infants (Rochat, 2013) and non-linguistic, non-human animals (Davidson, 1982; Stephan, 1999). Access to this stored conceptual information, in conjunction with several other cognitive capacities, allows us to identify new occurrences of these objects, whether it be in the perceptual identification of an object, in the physical environment, or the linguistic identification of an object referred to in discourse; where the object need not be present in the environmental context of the speech act.

In addition to these concrete nouns, we have abstract nouns. These nouns typically name ideas, states or qualities. They differ from their concrete counterparts in that they do not have a clear third-point physical object as reference, and are therefore not susceptible to the same form of PLT as described above. The absence of a third point implies that the conceptual content underpinning these types of nouns is acquired through point-to-point (P2P) interactions between two individuals, with one giving the information and one receiving the information. Continuing the teacher/child example, in this instance, the teacher would introduce an abstract noun such as 'politeness', 'truth', 'justice', etc… by demonstrating usage of this word to a child, giving the child the information that the teacher has stored in his/her brain in a direct P2P transfer, where the information transferred constitutes the stored conceptual information in the receiver. Importantly, this P2P information can be supplemented by a form of triangulation involving two or more individuals and a reference to a behaviour, let us call this: 'perceptuo-linguistic triangulation of behaviour' (PLTb). For example, a teacher, in further explicating the abstract noun: 'politeness', might draw the attention of child1 to the behaviour of child2 and say: 'that is how to be polite', where child2 is observed saying 'Please' and 'Thank you' in some social interaction. Here, the use of the adjective 'polite' is predicated onto child2's behaviour, where this behaviour constitutes the third point reference of the triangulation between the teacher and child1. With respect to the perceptual identification of abstract nouns in the environment, this is only possible with PLTb, because in P2P there is no

clear third-point physical object. Despite the differences between these two types of nouns, abstract nouns are used in discourse and theorising in much the same way as their concrete counterparts, in that they can occupy the same grammatical positions.

In addition to these communicative processes between individuals (PLT, P2P and PLTb), one may also receive conceptual information, relating to both types of nouns, from some other communicative mechanism, as opposed to a secondary person (the teacher, in the above). For example, an individual may learn about a noun's meaning from some form of media outlet, or from reading a certain branch of academic literature with its idiosyncratic and paradigmatic vocabulary (Kuhn, 1970), or perhaps even from the reading of a standard dictionary. In all of these cases, the mechanism is merely a deferred instance of an intentional human communication.

2.3 How is meaning structured?

Individual nouns, then, are associated with information. In philosophy, the long-dominant explanation of how this information is structured is referred to as the 'classical theory'. It is referred-to, as such, because of its dominance as a methodological approach in the philosophical discipline and because of its origins in the classical works of Greek antiquity, with Aristotle (384-322BCE/1933). Though, note that similar systems were popular in eastern cultures also, see Chakrabarti (1975) for a survey of the Sanskritic tradition of India, as one example.

A brief and critical account of this classical theory is advanced in the following pages, but prior to this, it should be acknowledged that this account is vulnerable to criticism at various points, for two reasons. Firstly, the account given is concise, and as such, by default, involves theoretical oversimplifications. Secondly, this account exaggerates the presence of the methodologically-pure version of this theory in the philosophical tradition. The reality of the matter is that the philosophers who operate within this framework (consciously or unconsciously) use the methodology in a much more nuanced way than described below. Nevertheless, it will be argued that the basic principles outlined do represent the implicit or explicit background assumptions of many philosophical positions, as well as positions developed in other disciplines. Finally, note that the classical theorist typically uses the term 'concept', when theorising about the structure and content of the information associated with words/nouns, but the meaning of this term is itself abstract, ambiguous and theory-laden. Therefore, the use of the term 'concept' has been minimised – though not in tracking the

historical accounts – favouring instead, fuller sentential descriptions to convey intended meanings. That being said, the phrases: 'word(s)', 'word meaning', 'lexical content/items' and 'semantic information/schematic' are used interchangeably to refer to the information underpinning nouns. These phrasal options aid the flow of writing.

When Ancient philosophers began to consider their experience of things in the world, it was quickly established that data acquired through the senses was not an ideal foundation upon which to build generalisable knowledge. Sense data involves the perception of individual and transient instances of things, which are prone to mislead e.g. the false perceptual bending of a stick when submerged underwater. Therefore, in order to achieve true knowledge, a theorist needs to abstract away from the sense data and work only with stable, idealised and universal conceptions of things, see Plato's (429-347BCE/1997): *Euthyphro*, *Lysis*, *Laches*, *Theatetus* and *The Republic*. In Aristotle's *Metaphysics* (384-322BCE/1933), we see this idea developed into a systematic analytical approach. Importantly for present discussion, Aristotle distinguished between the essential properties of a thing (a concept) and its accidental properties (384-322BCE/1933, Book VII). For Aristotle, an essential property of a thing is one that it must have to be categorised as a member of a concept. Whereas, accidental properties, on the other hand, are those properties that happen to be associated with the thing, but play no role in determining its essential nature.

In more contemporary terms, the classical theory conceives of the advanced form of conceptual information – that named by a noun – as rigidly definitional in structure, comprising of simpler concepts, properties or conditions (terms used interchangeably for the purposes of this description) that are individually necessary and jointly sufficient for category membership. The collection of necessary properties constitutes the precise extension of a concept. Different instances of the same concept necessarily share an identical set of simpler concepts, the details of which are explicated through *a priori* deductive reasoning, often (and hereafter) referred to as: 'conceptual analysis'. So, to say that a thing, let's call it X, is a member of the concept Y, is to say that X possesses all of the necessary properties associated with Y. Related to this, a thing cannot simultaneously belong to a conceptual category and not belong to it, the law of contradiction, e.g. X cannot be a member of Y and not be a member of Y at the same time. In addition, a thing must either possess or not possess a certain property: the law of excluded middle e.g. X must possess property Z or X must not possess property Z; the possession of a property is an all-or-nothing matter. In this sense, properties are binary, and as such can only take one of two values, either a property belongs to a concept (binary state = 1), or it does not

(binary state = 0); these values can also be framed in Boolean terms: 'true' or 'false', respectively. Holding to these assumptions, it follows that concepts have clearly demarcated boundaries, and once a conceptual category is established, it serves to unambiguously divide the world into two sets: things that are members of the concept, and those that are not. Things that possess the necessary properties of a concept are considered fully-fledged members of that conceptual category, and because property ascription is binary, there can be no 'better' members of a conceptual category than others, all members have completely equal status (Taylor, 1995, pp. 23-24).

The appeal of this approach comes from its ability to offer an objective, stable and unified explanation of concept/noun acquisition, categorisation, conjunction, disjunction, compositionality, and propositional truth and falsity, as well as providing the epistemological foundations for the analytical conception of our logical reasoning capacity and accompanying classical set theory (Lakoff, 1987; Osherson & Smith, 1981). In more general terms, the classical theory assumes that there are valid answers to the big philosophical questions, and that these answers are accessed through the critical evaluation of arguments using conceptual analysis.

Unfortunately, however, in the twenty-five-hundred years (or so…) of substantial classical analysis on the philosophically important questions and concepts, few, if any, classical theories have hit upon widespread agreement. In fact, as the research output increases exponentially, with the professionalisation of philosophy and advances in technology (electronic journals, etc.), so too does the divergence and/or disagreement associated with these concepts, as demonstrated in Chapter 1. After this period of time, it could reasonably be expected that there would be a higher rate of converging opinion and success, that is to say, if this method were to be considered a valid approach to knowledge acquisition.

In counter to this, it could be stated that this argument holds classical theory to an unreasonably high standard. For, even in the hard sciences, such as physics, there is rarely universal agreement on the nature of things, yet, we do not see a critique of scientific methodology to the same extent. This response is problematic for two reasons. Firstly, whilst there are disagreements in science, these are not nearly as common as those present in philosophy, which are ubiquitous across all topics. Secondly, whilst there are disagreements in science on the fundamental nature of things, nevertheless, the theoretical approximations produced in science are applied in the world with a high degree of success. Consider the science of atomic structure, there is no 'one-theory' that is considered to give the complete account on

the nature of an atom. Yet, the theoretical approximations explicating the nature of atoms are used very successfully in applied technology, for instance, in the manipulation of the hydrogen atom in magnetic resonance imaging (McRobbie, Moore, Graves, & Prince, 2006); just to name one application of the tens of thousands available in the modern world. In response to this, it could be argued that philosophy has a different and perhaps more difficult subject matter than science and therefore should not be held to the same standards. This may well be true, but it may also be the case that there is a flaw in a certain type of philosophical theorising, which explicitly or implicitly, consciously or unconsciously, assumes and partakes in important aspects of the classical theory of concepts e.g. making essentialist truth claims about a concept, telling the world what a concept is and what it is not, thus assuming definitional structure and necessary and sufficient conditions. This latter explanation is the position argued for in this thesis, with specific reference to the concept of self. Note that the points highlighted in this explanation are crucial to the overall thesis argument and will reappear in different forms as the thesis progresses.

Classical theories of concepts, then, seem only to function successfully when operating in relative abstraction from one-another, for when all theories on a particular topic are considered as a complete set, from a holistic perspective, various contradictions occur; as discussed in §1.6. To be clear, the main point being emphasised here is that the classical approach to explaining concepts and their structure has been unsuccessful in all attempts. Yet, nevertheless, aspects of this approach are still assumed and utilised in a wide range of theories, including those outlined in Chapter 1; this point is further developed in Chapter 3.

As Wittgenstein (1953/2009) and others have pointed out, in all but the tautological cases, the classical method of conceptual analysis, in attributing rigid definitions, has failed even with the simplest of concepts. More specifically, the explanatory definitions that spring from the classical approach fail to predict the referential range of a noun's meaning that we see in ordinary language applications (Labov, 1973), in that, there are always outlier instances of a noun's use that contradict any proposed rigid definition.

To counter this criticism, one could argue that the notion of accidental properties could account for this referential range issue, but this gives rise to another, arguably more difficult, set of problems, namely those presented against Gallagher's (2013) pattern theory solution, in §1.6. This failure in referential range is approximately equivalent to saying that the classical theory is too rigid and therefore, cannot account for the vagueness and fluidity that we see in the structure associated with noun meanings, reflected in ordinary language noun applications.

In specifying a precise extension for a noun's meaning, in the form of necessary and sufficient conditions, the classical theories appear to miss out on valid instances of its application, suggesting that a noun-meaning's structural extension, is in fact, fundamentally imprecise.

The first evidence of conceptual imprecision in referential range, coinciding with the first visible (empirical) cracks to the classical view, appeared with the development of 'prototype theory' (Rosch, 1973; Rosch & Mervis, 1975) and 'exemplar theory' (Nosofsky, 1988, 1992; Nosofsky & Johansen, 2000). On these highly related accounts, sustained exposure to the world and to language does not lead to rigid definitional meanings (i.e. classical/analytical concepts), but to cognitively-embodied, similarity-based, prototypical structuring; organised around the Wittgensteinian point of family resemblance (1953/2009). Here, members associated with a particular word's meaning do not have essential things in common, a prerequisite for the use of conceptual analysis, but rather are connected through a network of similarities or resemblances.

On the prototype account, word meaning has a probabilistic structuring, containing a set of properties (physical and functional) that are individually weighted in the definition of the word in question, each word specifying graded properties that its members tend to possess, a numerical probability between 0 and 1. As an example, take the semantic information anchored to the noun 'cat', an empirical survey of its typical properties might include: [1] overall shape (cat shaped), [2] four legs, [3] furry body/coat, [4] pointy ears, [5] tail, [6] whiskers, [7] retractable claws, [8] guile, [9] meows, [10] purrs, [11] likes eating fish and so on… These individual properties are weighted differently in the prototypical definition of 'cat', some more heavily than others. Take property [9]: 'meowing', this is more likely to be exclusively and heavily associated with the word: 'cat' than is [8]: 'guile'; a property shared by several other animals. The asymmetric weighting of these properties would be reflected numerically with 'meowing' having a probabilistic representation larger than 'guile', and hence closer to one. Note that, with all physical objects, those named by concrete nouns, there is good evidence that overall shape (as in [1]) is the initial and most heavily-weighted property in categorisation: (Jones, Smith, & Landau, 1991; Landau, Smith, & Jones, 1998; Landau, Smith, & Jones, 1988; L. Smith, B., Jones, & Landau, 1992; L. B. Smith, Jones, & Landau, 1996).

On the prototype account, a physical object is categorised as a 'cat' when it possesses some permutation of the above properties, such that they collectively surpass a statistical threshold of similarity in the individual perceiver (Verheyen, Hampton, & Storms, 2010). As for the use of 'cat' in linguistic discourse, as it is used in this paper, where no physical cat need

be present during the speech act, the suggestion would be that the word 'cat' brings to the mind of the speaker/listener a stored meaning made up of the statistically salient, heavily-weighted properties, restricted by some general information capacity restriction (Miller, 1956). The specific set of properties elicited in the individual is based on that individual's prior experience of the concept in question and is therefore subject to variance across persons, but due to the process of perceptuo-linguistic triangulation (PLT), this variance is kept to a minimum (discussed in more detail below).

With the exemplar account, word meanings are not represented structurally, by statistically graded properties, but rather, as a set of previously experienced salient instances of a thing, stored in memory. Continuing with the previous example, it could be imagined (at a stretch!) that an individual has only ever experienced cats of the Siamese variety, and so, on the exemplar account, these particular yet differing instances of Siamese cats constitute that individual's entire conception of 'cat' (i.e. a semantic schematic containing the above properties, except: [3] furry body/coat). If, then, at a later date, this individual chance-happens to encounter a Persian or a British Shorthair, where property [3] is present, then it is likely that these cats will also pass the resemblance threshold, set by the stored (Siamese) exemplar(s); given that there is much in common between these breeds (i.e. [1-11], but not [3]). Note, however, that though this example serves to demonstrate the idea of the exemplar account in a simple way, in reality this example is not conceivable, because the word meaning associated with 'cat' is used in a diverse set of contexts, both literal and metaphorical, and as such, goes beyond the simplicity of the visual information alluded to above. With the use of 'cat' in linguistic discourse, on the exemplar account the suggestion would be that the term brings to the mind of the speaker/listener a word meaning that captures the properties of the stored exemplar(s), again, subject to a general information capacity and a PLT restricted individual variance.

The important point to emphasise here is that both of these theories reject the notion that the semantic information underpinning nouns possesses rigid definitional structure. This stance allows them to circumvent the aforementioned criticisms levelled at the classical approach. Specifically, its inability to cover the full referential range of word's application that we see in ordinary language. Since both prototype theory and exemplar theory (hereafter, referred to collectively as: 'similarity-based theories') are working with forms of statistical similarity thresholds, rather than rigid definitions, they are not embarrassed by the lack of satisfactory definitions produced. As a demonstration of this point, consider a hypothetical

classical explanation/analysis of the semantic information anchored to the word 'cat', and for argument's sake, let the properties outlined above, [2-11], represent the outcome of conceptual analysis. These are the individually necessary and jointly sufficient conditions for the definition of an object to be categorised as a 'cat'; however, please note that [1] has been excluded, because it is difficult to conceive of how 'shape' can fit with the aforementioned classical assumptions, which is problematic in and of itself. However, putting this aside for the moment, *prima facie*, this definition seems reasonable, in that, most cats will have these properties. However, when we consider, more closely, the full referential use of the term in ordinary language, we see that it is applied to a variety of unusual breeds, some of which negate the aforementioned necessary properties. For example, there is the 'Devon Rex' which sports a soft coat (not furry [3]), the 'Munchkin' which has extremely short legs (reduced guile [8]), the 'Scottish Fold' with their ears folded forward (not pointy ears [4]) and the 'Ukrainian Levkoy' both hairless and folded ears (not furry + not pointy ears [3-4]).

The consideration of these outlier breeds testifies to the failure of rigid definitions; their inability to capture the full referential range of a word's meaning. Related to this point, Taylor (1995, p. 53) points out that, with only Aristotelian concepts at our disposal, new and novel data (or new and novel cat properties, in the above example) would require the creation of new words. Given that new data flows to the individual in a constant and rapid fashion, it would be highly inefficient if our semantic information were structured, as classical theorists would have it. In contrast, similarity-based theories can easily account for these non-typical applications, by postulating that the structure of 'cat' – and other word meanings, more generally – involves some non-fixed permutation of statistically salient properties. If an object possesses a sufficient amount of the properties associated with a stored meaning, then it will pass the similarity threshold for inclusion in that conceptual category. Also, with similarity-based theories, new data does not require the unnecessary creation of new words and meanings, nor a fundamental restructuring of the current words, for new data is typically integrated into existing semantic structures in a piece-meal, gradient fashion.

Another important advantage of these statistical models and all subsequent modifications of these perspectives, that toe the non-definitional, similarity-based, statistical line, is that, unlike the classical theory, they have been mathematically formulated and empirically tested, with much success: Fuzzy logic and fuzzy concepts: (Zadeh, 1965, 2015), Prototype theory: (Hampton, 1995; Rosch, 1973, 1975a, 1975b, 1977, 1999; Rosch & Mervis, 1975; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976), Exemplar theory: (Nosofsky,

1988, 1992; Nosofsky & Johansen, 2000; Storms, De Boeck, & Ruts, 2000), Colour terms: (Kay & McDaniel, 1978), Shape bias: (Jones et al., 1991; Landau et al., 1998; Landau et al., 1988; L. Smith, B. et al., 1992; L. B. Smith et al., 1996), Physiological accounts of category learning: (Shohamy, Myers, Kalanithi, & Gluck, 2008), Advanced mathematical models, quantum and other: (Aerts, 2014; Aerts, Broekaert, Gabora, & Sozzo, 2016; Verheyen et al., 2010). However, one caveat to this, it is not the case that these empirical results speak directly on the nature of the cognitive source mechanism/representation, that it functions in some similarity-based fashion, because the exact nature of the source mechanism/representation cannot be established through an examination of the structure present in the behavioural output. To proceed in this manner would constitute an unjustified form of reverse engineering or reverse causality, a *post hoc* fallacy such as: If A then B, B, Therefore A. Saturating this faulty logic with ordinary language we arrive at the following: (A) If we possess a similarity-based cognitive mechanism then, (B) this will be reflected and be detectable in our behavioural output, (B) our behavioural output reveals similarity-based structuring, Therefore, (A) we possess a similarity-based cognitive mechanism. Nevertheless, theorising of word/noun meanings (concepts) and cognitive processes should be mindfully constrained by these results and be able to give a strong account of why they occur, postulating possible mechanisms capable of producing such output, and finding ways to validly discriminate between competing mechanisms.

One final problem worth mentioning, with respect to the classical theory, is that even if definitions and definitional structure do/does exist, this still cannot be the complete story, because the definitions themselves need to be composed of some primitives that themselves would need explaining. As Fodor (1998, p. 44) points out, if some complex concept C is composed of the necessary and jointly sufficient conditions: C_1, C_2, C_3, $C_{N...}$, then to explain how we acquire and apply C also requires explaining how we acquire and apply C_1, C_2, C_3, $C_{N...}$and so on *ad infinitum*.

Given all of the above, it is difficult to envisage how the classical theory of conceptual structuring can hold, or can be justified as such, in light of the realities of ordinary language referential applications and modern psychological, mathematical and physiological evidence. This is not surprising, as the classical theory was not established on empirical or ordinary language considerations.

As discussed above, one of the fundamental problems with the classical conception is that it is too rigid. Postulating a core set of properties is too restrictive to the referential range

of a word's meaning and application. Similarity-based theories have the necessary degree of flexibility unavailable to the classical theory, and as such are able to account for the full referential range-of-use we see in ordinary language. However, in order for word meaning to function efficiently, it is imperative that it not be just flexible, but also stable. That is, in order that a word meaning and referential application not become chaotically random, it should possess an in-built propensity towards structural stability. This stability is a balancing act between not allowing drastic change with the arrival of new data, but simultaneously being flexible enough to adapt itself to changes in use and context, brought about by social forces. On the classical account, stability arises with ease, from the perfect correlation of properties over a word's members, but this stability comes at the cost of rigidity in referential range. In similarity-based theories, the heavily weighted features of a concrete noun most certainly approach classical core status, but stop just short of the classical rigidity. The next important question is 'how is stability possible?' It is argued, hereafter, that we need triangulation for this stability to occur.

On the similarity-based accounts, with reference to the perceptual identification of concrete nouns, it is the case that the semantic information associated with these nouns have relatively clear and relatively stable inclusion/exclusion criteria, for whether an object is a member of the word category or not. This criterion is justifiably considered 'relatively clear', because when we look at the application of concrete nouns in ordinary language, we rarely see problems in discrimination. For example, any linguistically competent, neurotypical adult can easily distinguish between a 'cat' and a 'car', and it is unlikely that an argument would breakout between individuals over this sort of distinction. This relative clarity has arisen, because the objects named by the concrete nouns have a shared existence in the perceptual domain. It is in this shared perceptual domain where these concrete nouns are frequently exposed to PLT, and it is in the PLT process where the consensus driven boundaries of a word meaning are formulated. By analogy, in the same way that the electrical force (more-or-less) keeps the atoms of an object together, including its boundaries (though, boundaries are more susceptible to change), so too PLT (more-or-less) keeps the boundaries of a word meaning, clear and stable.

The constant stream of PLT references to objects in the environment incrementally strengthens the sensitivity of our semantic schematics, our word meanings. Besides this, the stability is also a consequence of the fact that word usage (the 'L' of PLT) rarely undergoes radical change, though, of course, it does evolve over time. Take, for instance, the use of the word 'car', initially applied to the 1885 Benz Patent Motor Car, but now applied to Google's

Silicon Valley driverless cars. Here, from the term's inception through to its modern incarnations, we see a significant change in usage, but a clear hereditary resemblance remains. On this point, there are strong (but still largely conjectural) arguments, orientated around adaptive pressures, for why an element of vagueness in language and word meanings is necessary. Meanings, named by concrete nouns, used in a complex world with constantly-updating pragmatic context, have to be imprecise in order that referential communication be successful (Hampton, 2011). PLT, as a mechanism, permits this form of change, because when word usage changes, so too does the PLT process. As for the use of concrete nouns in linguistic discourse, where the object in question need not be present, the stability afforded by PLT means that the uttering of a concrete noun elicits an approximately equivalent set of properties in both the speaker and listener and these properties are likely the heavily weighted properties, constrained by some general information capacity restriction (Miller, 1956).

This proposition, that only the heavily weighted properties of word meanings are recruited during language processes, is indirectly supported by the empirical literature on semantic illusions. For example, in Erickson and Mattson's (1981) 'Moses illusion', participants are asked: 'How many animals of each sort did Moses put on the ark?'. Erickson and Mattson found that participants typically state 'two', without registering that Moses was not the biblical character associated with the ark story, Noah was. Similar results have been obtained with questions such as 'After an air-crash, where should the survivors be buried?' (Barton & Sanford, 1993) and 'Can a man marry his widow's sister?' (Sanford, 2002). In these cases, listeners seem only to be processing a restricted and context sensitive set of properties associated with a particular word meaning, only shallow dipping into the word's meaning. This shallow-dipping approach is most likely a consequence of efficiency pressures, i.e. the need to rapidly process large amounts of incoming linguistic information. Nevertheless, because of PLT, the set of properties elicited by concrete nouns is approximately equivalent between speaker and listener.

As for the similarity-based account of abstract nouns, those that largely derive their content from the exchange of information between individuals (e.g. P2P), these terms, devoid of a clear third-point-physical-object, are not susceptible to PLT. As such, their inclusion/exclusion criteria are much fuzzier in comparison to concrete nouns. Rosch and Mervis (1975, p. 576), on this point, state that such categories 'are of particular interest [from an empirical perspective] because they are sufficiently abstract that they have few, if any, attributes in common to all members'. The likely reason for this fuzziness is that, unlike

concrete nouns, abstract nouns do not relate to objects in the shared perceptual domain, and are therefore not subjected to perpetual PLT, where the consensus driven boundaries of a word meaning are formulated and stabilised. If you define a word in terms of the content of an individual mental state, as occurs with P2P interactions, then the word just means whatever is in that person's mind. Not unexpectedly, then, discussion of what constitutes properties and examples of abstract nouns can be problematic and prone to severe disagreement, especially when subjected to theoretical scrutiny, as in conceptual analysis; as in Chapter 1; see also Gettier's (1963) account of 'knowledge' as 'justified true belief' and the numerous responses to it, as the classic, controversial, philosophical case in point; see also Wittgenstein's (1953/2009) discussion of games.

These types of disagreements also occur in everyday conversations. For example, in discussions on who is the best: athlete, sports team, singer, actor, etc. debate often arises because individuals have different meanings of what constitutes 'best', 'skill', 'excellence' or some other related word. Consequently, and unknowingly, people involved these debates are often talking at cross-purposes. This is true for those with a 'loftier' taste in conversation; take Duchamp's 'Fountain', the signed urinal art-exhibit. This piece strongly polarises opinion, because of the differences between individuals of what the abstract noun 'art' means.

In short, the meanings associated with abstract nouns, processed through P2P interactions, possess high degrees of variance with respect to their meanings and their inclusion/exclusion criteria, whereas concrete nouns, processed through PLT, have relatively clear and stable meaning and criteria. An important modification to abstract nouns, relating to perceptuo-linguistic triangulation of behaviour (PLTb), will be elucidated in Chapter 4.

Osherson and Smith (1981) are one of the exceptions to this growing consensus towards similarity-based theories, as advancing the correct characterisation of semantic information structure, underpinning nouns. They maintain that word meanings are structured in the classical sense, as a core set of necessary and sufficient properties, and that these define the real essence of a category. Whereas, factors associated with statistical similarity-based theories function as part of identification procedures, only: 'the core is concerned with those aspects of a concept that explicate its relation to other concepts, and to thoughts, while the identification procedure specifies the kind of information used to make rapid decisions about membership' (Osherson & Smith, 1981, p. 57). The example they develop is on the concept 'women', were the classical core properties refer to information about the presence of the female reproductive system, while the identification procedures operate on properties such as

body shape, hair length, and voice pitch. Osherson and Smith use this distinction in order that they can combine the aforementioned theoretical advantages associated with the classical account of word meanings with the strong empirical evidence associated with the similarity-based theories.

The problem with assuming a form of duality to semantic information, as do Osherson and Smith, is that they are left with the difficult task of explaining how the tension between the two components is resolved. Specifically, explaining how they relate and interact with one another in cognition. On this point, given that the full spectrum of cognitive faculties (i.e. knowledge, memory, attention, belief, words, and their underlying semantic structure) are all intimately bound-up together and employed in these identification procedures and ordinary language applications, it is not clear what the justification is for postulating the existence of classical core properties, beyond these. Perhaps, besides holding to some strong metaphysical assumption that is likely immovable and/or unfalsifiable.

A more targeted criticism relates to the example they select to demonstrate this duality, specifically, the concept 'women'. It is not clear how Osherson and Smith account for the biological heterogeneity associated with the female reproductive system (i.e. individual physiological differences in the reproductive system, across women). Also, in the case of women who have had a partial or complete hysterectomy, are we to say that these are no longer women? This seems wrong-headed.

Osherson and Smith's main argument for this duality is that it facilitates certain theoretical advantages that cannot be achieved in similarity-based theories; particularly, prototype theory supplemented by fuzzy logic. These include the combination of concepts to form complex expressions and truth conditional propositions. Several mathematical and empirical papers have forwarded strong theories to refute this claim: (Aerts, 2014; Aerts et al., 2016; Zadeh, 2015), yet this criticism still lingers on, but now as part of a much wider point that any would-be theory of meaning needs to address, that is, the principle of compositionality.

2.4 Compositionality and its failure in meaning

A systematic examination of compositionality, with a critical eye towards similarity-based theories of meaning, came from Fodor (1998), and was supported later by Gleitman, Connolly, and Armstrong (2012), amongst others. Fodor argues that any theory of word meaning should satisfy certain non-negotiable conditions, one of which is the principle of compositionality. The principle of compositionality states that the meanings of a complex expression are

determined by the meanings of its parts, in conjunction with some syntactic rules that govern their combination. An explanation of compositionality is held as non-negotiable, partly because it soothes the productivity paradox; the perceived tension between the finite brain (its finite storage capacity) and the seemingly infinite possible language outputs (the potentially infinite class of meaningful sentences that can be produced and comprehended). In real terms, with reference to comprehension, the idea is that a speaker of a language is able to understand the meaning of any grammatical sentence uttered in that language, even if the speaker has never experienced the sentence before. In this sense, the individual (theoretically) knows infinitely different sentence meanings, despite the apparent finite nature of the brain storage capacity. Adopting the principle of compositionality, in which there are finitely many words in the lexicon (our 'mental' dictionary) and finitely many grammatical rules governing combination, is one way to efficiently resolve this tension, for infinite permutations are possible when these two finite aspects interact. The words provide the meaningful content and the syntax, free from meaning, combines these words in some ruled-based fashion.

It is Fodor's contention that statistical-similarity theories do not explain compositionality, because they cannot account for the manner in which the proposed prototypical structures (underpinning words) are combined. More specifically, Fodor states that the prototypes of complex concepts (multiple words or phrases) are not statistically related to their constituent prototypes in ways that would explain productivity (1998, p. 100). For example, in the complex expression 'pet fish', the two constituent words 'pet' and 'fish' are combined, but it is not clear that the meaning of the complex expression (it's prototypical properties) have a basis in the meanings of its constituents' 'pet' and 'fish'. In other words, the meaning of 'pet fish' is not achieved by binding the prototypical/statistical meanings associated with 'pet' and 'fish', for in their combination something different and statistically unpredictable emerges. This is an unusual example, for it could be argued that a pet fish (e.g. a guppy), is very close to fish and pet prototypes, thus, the emergence of guppy is not that surprising. More problematic than this, is the fact that the criticism assumes that because individual words appear to be structured and processed in a statistical fashion, it is therefore the job of syntax, in like manner, to combine these individual word meanings in a statistical way. In other words, it is the job of syntax to statistically combine concepts/word meanings, the sum of which is a complex meaning, a meaning which is unaffected by the syntactic computation.

As an alternative to this, it is possible that the mode of combination could itself be syntactic, that there is some non-statistical syntactic form to the combination of word meanings

that holds even if the input into the syntax is a statistically defined word or concept. Fodor's own perspective is based on a foundational idea that individual word meanings are the basic building blocks of complex meaning, semantic primitives which possess within their meaning some form of reference to a particular external object; presumably a non-linguistic referential element. On this view, syntax contributes nothing to meaning, it simply combines the lexical meaning there is; where the mode of combination is conceived as non-syntactic, thus it cannot even constrain meaning. However, there is no clear evidence to suggest that this is the case or why it need be so, and a closer examination of the compositionality framework, reveals that it is lacking in explanatory power with respect to important aspects of meaning namely, referential and propositional meaning (Hinzen, 2014; Hinzen & Sheehan, 2013). On this point, Hinzen and Sheehan argue that grammar has a much more influential and formative role in complex meaning than simply the mechanism utilised to concatenate concepts, as assumed on the compositionality account. Instead, grammar – the set of structural rules governing the construction of words, clauses and sentences in a given language, spanning the fields of morphology, phonology, prosody, syntax, semantics and pragmatics – gives rise to the forms of meanings most important to human functioning and communication.

Before further exploring these grammatical forms of meaning in detail, first an outline of where compositionality arguably fails, based roughly on the arguments developed by Hinzen and Sheehan (2013, pp. 88-95). Take the sentence 'Victoria sleeps'. On the standard compositional account, the meaning of the sentence 'Victoria sleeps' is composed from the independent word meanings assigned to 'Victoria' and 'sleeps', in conjunction with some syntactic operation.

One plausible account of this syntax is Merge; the basic operation from Chomsky's (1995/2014) minimalist program. Merge is hailed as the simplest procedure that can achieve recursion, which in turn, satisfies the principle of infinity discussed above. At bottom, the Merge procedure involves taking two linguistically formed units and merges them into a bigger unit or set. On this perspective, the formal generation of the sentence 'Victoria sleeps' will look as follows: Merge (Victoria, sleeps) = [Victoria, sleeps]. So far there is nothing in this conception that clarifies why the particular complex meaning that we comprehend actually arises, instead of some other meaning. Minimally, we need to project an ontology onto these words: of objects (e.g. Victoria) and properties/predicates (e.g. sleeps) and specify how they combine to give rise to referential and propositional meanings. The compositional semantic perspective does just that, in that, the ontology of objects and properties/predicates is taken to

be part of the lexical meanings: proper names contribute the objects, verbs the properties. This response is unsatisfactory, because referential and propositional meanings are never found in individual lexical items (argued below and in §3.3), and neither can individual lexical items, in isolation, be grammatical subjects or predicates. It is only in a grammatical context, in the relation between words, where these possibilities can arise. Stated in other words, it is only from within this grammatical system that we can initiate the referential and predicational process, the result of which is propositional meaning, that is, sentences which possess the possibility of truth and falsity (Hinzen, 2014, p. 235). The overarching idea of the compositional semantic perspective rules this possibility out, for it assumes that syntax/Merge plays a role independent of meaning, in the sense that it only combines meanings that already exist lexically. This prohibits the compositionality perspective from accounting for reference, for individual lexical meaning does not account for reference, and syntax, in the sense invoked in the compositionality principle, cannot provide it either. On the matter of propositions, a compositional semanticist might take this response a step further by arguing that there is a function lexically specified in the meaning of 'sleeps', which maps 'Victoria' onto the proposition that Victoria sleeps. But since the proposition in question is invoked in the definition of this very function, this doesn't explain how the function gives rise to the proposition (the explanation is circular). Therefore, the compositional account runs into three problems: (1) By depriving grammar of meaning, it forces the explanatory factors to be lexical. (2) But the kinds of meaning required (referential meaning and predication) are never available purely lexically (discussed in more detail below). (3) Their definition is circular. Hinzen & Sheehan conclude that we can only break this explanatory regress by endowing grammar with meaning: it engenders the formal-ontological distinction between object and properties required, through its distinction between referential and predicational expressions, which cannot be grounded lexically or compositionally.

Referential meaning is a specific form of meaning that involves a speaker picking out something in the external world, whether it is an object, action, or event. This form of meaning does not arise at the individual word level. For example, the word 'cat' in isolation, picks out a general concept/meaning, but it cannot, in isolation, pick out one cat over another, refer to the cat I see now or saw yesterday, or a particular group of cats. These are referential distinctions, which arise at the grammatical level; primarily, in the noun phrase (NP) grammatical configuration (explained below) embedded in a contextually appropriate sentence. It follows from the above that the introduction of grammar does not create more

words/concepts, but rather that stored words/concepts enter into this grammatical system, which converts them into referential meaning.

Another means of communicating referential meaning is via gesturing, which can also be argued to be grammatically complex. For example, if the word 'cat' is accompanied with some form of pointing gesture, [pointing gesture + 'cat'], then the pointing gesture acts in much the same way as the determiner or demonstrative in an NP, and as such, reference becomes possible; in line with longitudinal correlations between index-finger pointing and determiners (Iverson & Goldin-Meadow, 2005). To be clear, the meaning associated with the word 'cat' cannot account for how the word functions grammatically, whether it is inflected for number, whether it is the complement of a determiner, whether it will place in a referential/subject or predicative position, whether it is a head, modifier or adjunct, and so on. Because of this fact, it cannot account for referentiality, no word considered in isolation can. This point also extends to determiners, demonstratives and indexicals, in isolation. 'a', 'the' and 'this' (also, 'I' and 's/he') do not have a referential meaning when considered as individual items (a point revisited in §4.3), neither do they have lexical content, but they do indicate grammatical relations at the morpho-lexical level, in its regulation of reference and propositional meaning.

One obvious objection to this position needs addressing. This relates to the referential aspects of proper names. The objection would be that proper names, such as 'Victoria', considered as individual words, seemingly devoid of grammatical features and relations, function referentially in a uniform lexical fashion. Intuitively, this sounds plausible, but in reality, proper names can and do function predicatively in many sentences, and if a proper name does function referentially, it is only when occurring in the appropriate grammatical configuration. For example, consider the sentences:

(1) Peter is a student in my reading group.
(2) There is no Peter in my tutorial group.
(3) The Peter I am talking about is not the Peter you know.

In (1), 'Peter' is functioning referentially in subject position, with a definite sense that a particular individual from a reading group is referred to, whereas in (2), 'Peter' is functioning predicatively, as a general description of people named 'Peter'. Here, and in general, how a proper name refers precisely co-varies with the grammatical role that it plays. A proper name

functions referentially when it appears grammatically in a referential position, and functions as a predicate when it appears grammatically in predicative position (like as the restrictor of the quantifier 'no' as in (2) or the complement of the determiner 'the' in (3), as any common noun normally would).

One counter response to this is that all proper names, as members of the noun word class, function in much the same way as other nouns in this class, namely, as the predicative complement of a determiner, but in the case of proper names the determiner is removed and unpronounced (Fara, 2015). In response to this, Hinzen (2016) outlines nine problematic consequences of adopting this position, the conclusion of which is much in keeping with the perspective developed above: that there is a strong grammar/meaning alignment that undermines any motivation to posit the existence of covert determiner syntax in this particular case. In short, the meaning or lexical content associated with words, considered in isolation, has no control over the form of reference it will be used in.

Referential meaning depends on forming NP configurations that (sometimes) include a determiner, which is a function word devoid of lexical content, but serves to indicate grammatical structuring, and the noun, which has lexical content associated with it, as discussed in §2.2 and 2.3. Finally, you have the grammatical NP structure itself that the determiner and noun enter. Full referential meaning is achieved only when this NP configuration is in turn embedded in a contextually appropriate sentence, which is fully saturated with information from all branches of grammar, plus the relevant non-linguistic information and presuppositions of the speaker using it.

It is not only referentiality that depends on grammatical structuring, but also the exact form of reference that is executed depends on the particular grammatical configuration generated, specifically whether it be *abstract/generic, indefinite, quantificational, definite, rigid, deictic or personal*. In fact, there is a positive correlation between grammatical complexity and the definiteness of reference. Cross-linguistic evidence suggests that as grammatical complexity increases (from left to right, in the above italicised list) so too does the definiteness of the reference (Hinzen, 2014; McNally & Van, 1998). In English, we observe that at the weak end, there are abstract/generic and indefinite NPs, which take no determiners or weak determiners (e.g. '**Men** like cheese', 'I ate **a chicken**'). Whereas, at the strong end, there are definite and deictic NP configurations that obligatorily take (strong) determiners, omission of which isn't possible without altering meaning (e.g. '**the/this man** fell over); for further explication of the strong-weak distinction see McNally and Van (1998). As reference

becomes deictic, further functional morphemes are added (e.g. 'this', which decomposes as 'the+here'), and finally grammatical person distinctions are added, which play no role in lexical nominal (this hierarchical grammatical structuring is full detailed in §3.3).

In summary so far, the argument of this chapter has unfolded as follows: First, the differences between concrete and abstract nouns were highlighted. Concrete nouns name things that have physical properties and are therefore susceptible to PLT. Conversely, the things named by abstract nouns do not have physical properties and are therefore not susceptible to PLT; but they can be linked to PLTb (more this idea in Chapter 4). On this point, it was argued that it is through these triangular processes that word meanings become sufficiently stable to be used successfully in acts of communication. Next, it was argued that word meanings are not definitions, in that they are not structured with individually necessary and jointly sufficient components. Instead, it is highly likely – based on theoretical and empirical evidence – that they are structured via some similarity-based, statistical system. Unfortunately, these similarity-based explanations of word meanings do not easily satisfy Fodor's (non-negotiable) compositionally principle: that meanings of a complex expression are determined by the meanings of its parts, in conjunction with some syntactic rules that govern their combination. In response to this, it was argued that compositionality is the wrong principle to account for complex meaning beyond individual words (i.e. grammatical meaning). Grammar itself has to be factored into the equation that generates grammatical-level meaning from individual word meaning. It is the cognitive principle needed, distinct from statistical learning of lexical features, that allows word meaning to function referentially.

2.5 A computational parsing model of grammar and referential meaning

On the Hinzen & Sheehan model, referential meaning occurs only in grammatical NP structure, which, by necessity, must be embedded in a contextually appropriate sentence. This grammatical NP structure is not sensitive to the internal feature composition of word meanings but is sensitive to the grammatical properties they possess. For example, the configuration: [$_{NP}$ [D] [N]] facilitates the possibility of a particular form of referential meaning. The computational process doesn't care about the particular noun-meaning [N] entering into the structure, whether it is 'cat', 'cow' or 'comb' etc… However, it will be sensitive to the grammatical aspects of these terms, specifically, whether the form is singular or plural, indicated in the suffix by adding -es or –s (in English) and it will also be sensitive to the

grammatical category of the word. The same is true of the determiners [D]. Although determiners are devoid of lexical content, they do indicate the typical form of grammatical structuring that will follow. But how exactly does this philosophical/grammatical explanation of meaning fit into fully explicated formal theories of grammar and computational explanations of parsing? On this matter, there is one particularly influential computational account of grammatical parsing, developed by Vosse and Kempen (2000), which will play an important explanatory role in Chapter 3, 4 and 5. For the purposes of the present chapter, this parsing model will be considered as a plausible candidate mechanism for the production and comprehension of referential meaning as broadly outlined above, as a grammatical phenomenon. Besides the explanatory role that this model plays in this thesis, this model also accounts for a variety of language processing effects, namely, lexical syntactic ambiguity (detailed below), as well as explaining the behavioural comprehension performance of aphasic patients, a factor barely considered in theories of parsing and grammar more generally. This model has been further extended by Hagoort (2003a) and others, to account for neurophysiological neuroimaging evidence (see below and §5.4). Since its initial publication, seventeen years ago, this theory is still cited as a leading explanation of grammatical processing (Hagoort, 2017).

The general architecture of the Vosse and Kempen model, hereafter termed the Unification Model (u-model), is based on the foundational view that each word or lexical item that exists in a person's lexicon is linked/stored with a grammatical, structural frame. It is Vosse and Kempen's (p. 110) contention that this structure is in the form of a three-tiered hierarchically connected tree (or four-tiered, if you count the lexical item as a tier, as is assumed hereafter), with the top layer consisting of a single root node (e.g. Sentence (S), noun phrase (NP), preposition phrase (PP), determiner phrase (DP)). The second tier, connected to the root node, consisting of one or more functional nodes (e.g. subject, head, direct object and modifier); importantly, besides modifiers, only one instance of each functional node is allowed in the same structural frame. These functional nodes (in tier 2) are connected to, and dominate, a third tier phrasal node (e.g. Determiner (D), Noun (N), Verb (V) Preposition (Prep), Adjective (Adj) and Adverb (Adv)) to which the individual lexical items (words) are attached; the fourth tier – see Figure 1 below.

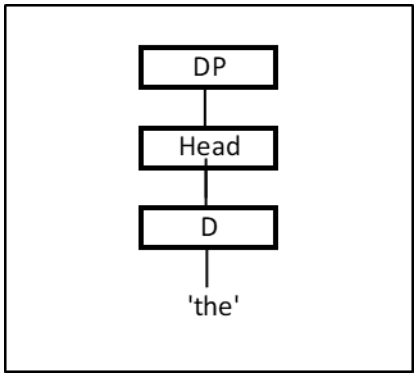

Figure 1: Grammatical, structural frame associated with a word. A modified figure from Vosse and Kempen (2000, p. 110).

What is unique about this account is that all of the grammatical structure associated with the root notes are stored (in chunk form) and are retrieved from the lexicon when words are selected, such that, words prime the unfolding of specific grammatical, structural frames. This perspective almost completely dissolves the lexical-syntax distinction, in that, syntactic structure (one aspect of grammar), rather than being something separate from the lexicon, is stored and utilised as part of lexical encoding and retrieval mechanisms, respectively. Hence, there are no additional, lexically-dissociable, formal grammatical rules that are introduced beyond those associated with the lexical item itself. Stated differently, there are no independent syntactic rules that introduce additional nodes.

In the language comprehension process, the grammatical, structural frames associated with and elicited by individual words, when they are read or heard, enter into what Vosse and Kempen call a: 'unification work space', hereafter 'u-space'. As the newly inputted words enter this u-space, parse-tree structure assembly begins. Specifically, the structure associated with the incoming, individual lexical items links-up together through an operation process, which connects identically labelled root and foot nodes from the individual structures, to bridge the full length of an utterance; the link between structures is hereafter referred to as 'u-link(s)'. There are specific preconditions for structural assembly of u-links to occur, and these are that the grammatical features, of the individual structures to be linked, are compatible. This involves the checking of grammatical agreement features, such as: number, gender, tense, case and person, as-well-as word order checks. In the Vosse and Kempen example sentence: 'The women sees the man with the binoculars', (2000, p. 110) – which is deliberately ambiguous, a facet that will be ignored for current purposes – the first three words that enter the u-space are: 'The', 'women' and 'sees' as depicted in Figure 2, below.

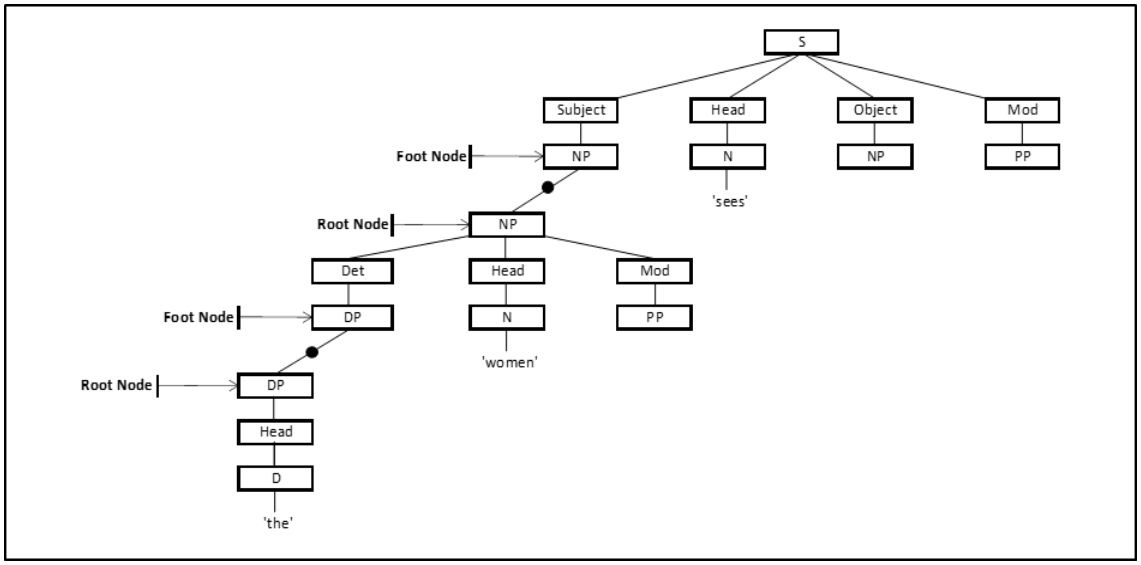

Figure 2: Reworked and extended example of the Vosse and Kempen (2000) unification process. This figure depicts the unification operation of three lexically-specified syntactic frames. The first unification (or u-link) links the root node DP (associated with 'the') to the foot node of the same available DP category slot (associated with 'women'). The second u-link links the root node NP (associated with 'women') to the foot node of the same available NP category slot (associated with 'sees')

When the determiner 'the', in Figure 2, enters into the u-space, it primes a certain grammatical, structural frame that restricts the possible words and frames that can follow, based on the specifics of the associated structure it elicits. Following this, 'women' enters the u-space, at which point a u-link is proposed between 'the' and 'women', for there is a mutually compatible unfilled slot between the already existing structural DP frame, elicited by 'the' – which allows for a noun as a complement to the determiner – and the empty determiner slot introduced by the structural frame associated with the noun 'women', which is open to the possibility of a determiner filling this grammatical slot. Next, the verb 'sees' enters into the u-space, bringing with it, its own grammatical structure. Part of the verb 'sees' grammatical structure is the subject NP slot, which carries the following grammatical features: case is nominative, person is 3rd, and number is singular. These are compatible with the root N node 'women' of the NP 'the women'. Specifically, 'woman' carries the following grammatical features: case is nominative or accusative, person is 3rd and number is singular. Hence, the grammatical agreement feature check is satisfied. The second check relates to word ordering constraints, in that, the root node of a lexical frame is only allowed to unify with the foot node of another frame if this does not violate the precedence rules for the branch dominating the foot note (see

Vosse & Kempen, 2000, p. 111; for full explication of precedence rules). If both of these grammatical checks are satisfied, as they are in the above example, the u-link is completed.

The u-links unified in this u-space are formed dynamically so that the strength of the u-link fluctuates continuously until a stable resting state is reached, typically where the utterance meaning is fully established. Given the unpredictability of natural language, from the infinite combinations that are technically possible, it is the case that several alternative binding candidates will exist and be available (to bind) at every stage of the unification operation. From these competing candidates, one configuration will typically arise triumphant. In order for this to occur, it is necessary that one binding candidate remain active throughout, with equilibrium reached through the lateral inhibition (or competitive inhibition) of the alternative u-links. Lateral inhibition is the force that Vosse and Kempen invoke to account for the changes in u-link strengths. When a frame gives rise to two or more possible u-links, these possibilities act to mutually inhibit one another by emitting an inhibitory force on all competitors, reducing the effect of strength of competing u-links. How this works is that every viable u-link is associated with a numeric value between 0 and 1 which expresses the strength of the binding between the linking nodes; the root and foot nodes of individual lexical structures. When a u-link is initiated, its initial value will be at 0 (poor goodness of fit) and in the absence of counteracting, competitive u-link candidates, with their inhibitory forces, this value will quickly rise to 1, or thereabouts ('1' being equivalent to perfect goodness of fit), at which point the u-link can be said to be attached. With respect to local competitions, the system selects the strongest u-link for every root node that lands on the node in question, providing its value is > 0.50 and it satisfies the grammatical preconditions outlined above. A parse is successful if one and only one u-link is selected, meaning that all other competitive u-links will fall by the wayside as their value attenuates towards 0. At the end of this successful parse there should exist a connected configuration of unified root-foot pairs, which comprises of one grammatical, structural frame for every word inputted, except for the upper-most sentence node.

Besides the attenuation in unsuccessful competitive u-links, there is also a more general attenuation in activation in the early units of the utterance, with the most recent nodes having a higher level of activation in the u-space, in comparison to the earlier nodes. Consequently, the former will have more inhibitory leverage than the latter. Furthermore, the strength of the u-links is also a function of semantic plausibility (e.g. thematic and pragmatic fit). For example, semantically similar words tend to co-occur more frequently than unrelated words; this is basis of lexical semantic analysis (Landauer, Foltz, & Laham, 1998). These

semantic factors can override the attenuated activation effect and prime certain u-links over others, but, by their own admission, Vosse and Kempen have not factored-in the role of word and referential meaning in this computational parsing theory. This is something briefly touched-upon below; with reference to the grammatical meaning discussed above, but first, a summary of the u-model.

There are six key aspects to this Vosse and Kempen model (2000, p. 136): [1] This computational parsing system assumes that grammar is fully lexicalised, in the specific sense that chunks of grammatical structure are stored and retrieved from an individual's lexicon. In simpler terms, grammatical structure is stored in (or associated with) individual words and is activated using these words. These chunks are the elementary building blocks, from which, fully grammaticalised structures are built. [2] There is a unification operation that combines the grammatical, structural frames associated with the individual, incoming words, in to larger structural frames that span the full length of the utterance. [3] The unification operations that take place are all local, in the sense that it is a linear process with locality set by the unfolding structural frame. [4] All u-link attachments are graded as opposed to binary. These u-links fluctuate between 0 and 1 as different inhibitory forces act upon them. [5] These inhibitory forces are the product of mutually exclusive attachments, which compete with one another continuously. The dynamic mechanism responsible for the competition is referred to as lateral inhibition and this controls the u-link attachment strengths and consequently the parse-tree configurations that unfold. [6] Finally, there is a decay value that occurs as processing unfolds. This affects the inhibitory power of all units that enter the u-space, but especially those that occur during the early stages of an utterance.

Hereafter, it is speculatively suggested that the u-model may also account for the specific forms of referential meaning that begin to take form in grammatical configurations. Specifically, the different chunks of structure stored and associated with particular lexical items, may correlate with or constrain particular forms of referential meaning that eventually unfold. The structural frames elicited by individual words, when fully explicit in phrasal form – as part of a contextually appropriate sentence – can impress upon their constituent units a certain form of referential meaning, a referential type, and this referential meaning/type will be blind to the lexical content of the word/noun that enters it. For example, the determiner 'the' typically (but not always) indicates the unfolding of a definite NP configuration (e.g. [$_{NP}$ [D] [NP]], as in 'the women'); see §3.3 for a full description of NP types and their relation to referential meaning. Therefore, it is possible that the determiner 'the' primes a structural frame

to accommodate the likely unfolding definite NP configuration. These particular definite NP configurations restrict reference to a particular thing (or things) in the world, in the definite sense. Therefore, it is possible that this stored frame, besides containing structural information on grammar, also primes a certain form of referential meaning upon the constituents that enter it, namely, a referential meaning to a thing (or things) in the world. This point may also hold for other demonstratives/determiners such as 'this' and 'a', which (respectively) are typically associated with a deictic NP configuration (to a specific physical thing in the here-and-now environment of the utterance) and to an indefinite NP configuration (an indefinite reference to a thing (or things) in the world). Therefore, like 'the', it is possible that the structural frames associated with 'this' and 'a', and any other demonstratives/determiners, prime and impress their own individual forms of referential meaning upon the constituents that enter their structure.

An infinite number of computational models could be developed to explain the linguistic behaviour of humans. Thus, discriminating between these (i.e. identifying the good/useful/true models from the bad ones) is no mean feat. This process of identification is aided by considering how a computational model could be instantiated in the hardware it has arisen from, for the nature of the hardware can provide important constraints on modelling possibilities. For instance, the concrete, observable, nature of hardware, facilitates the possibility of engaging the scientific method, in particular, the formulation of clear scientific predictions, which any computational model would have to predict and explain. For current purposes, with respect to the Vosse and Kempen (2000) model, this would require answering the question: 'how is this u-model instantiated in the brain?' This question in particular, because the consensus view is that the brain, as part of the central nervous system, is the hardware (or wetware) responsible for human cognition, including the language faculty. This is an area that Hagoort (2003a) and others have elaborated on, exploring the u-model in light of the neurophysiological data on language processing. Hagoort argued that there are two key areas that are critical for the form of grammatical processing described by Vosse and Kempen: the left posterior superior temporal gyrus and the left inferior frontal cortex. This finding was initially based on a meta-analysis of 28 neuroimaging studies performed by Indefrey (2004) and was supported by later by Snijders et al. (2009). The main idea behind this theory is that the left posterior superior temporal gyrus, known to be involved in the processing of individual lexical items (Hickok & Poeppel, 2004, 2007; Indefrey & Cutler, 2004), could also constitute the section of the brain that is vital for the storage and retrieval of grammatical frames; frames

stored as part of the lexicon, associated and activated by individual words, and potentially associated with forms of referential meaning, as alluded to above. In addition to this, we know from research examining brain lesions in the left inferior frontal cortex – as experienced in certain forms of aphasia – result in deficits in syntactic processing (Caplan, Michaud, Hufford, & Makris, 2016). Based on this evidence, Hagoort argues that this might be the area where the individual grammatical frames are connected into larger phrasal and sentential configurations, the u-space where u-links are formed. In short, the hypothesis is that the left superior posterior temporal areas of the brain are where grammatical frames are stored and retrieved, and the left inferior frontal areas of the brain are where the grammatical frames are connected with one another to form larger phrases and sentence. Further evidence for this neural model will be discussed in §5.4, with reference to electroencephalography (EEG) neuroimaging research.

2.6 Chapter conclusion

The argument of this chapter has covered a wide range of ideas. First, it was argued that statistical, similarity-based accounts of meaning structure, better explain the empirical data and theoretical necessities associated with word meaning, than the classical/analytical accounts. Further to this, a differentiation was made between the nature of concrete nouns and that of abstract nouns. Specifically, it was argued that concrete nouns name things that have a shared existence in the perceptual domain, thus, by default, possess physical properties susceptible to PLT. In contrast to this, abstract nouns name ideas, states or qualities, which do not have physical properties and are therefore not susceptible to PLT (however, they can be linked to PLTb, more on this form of triangulation in Chapter 4). It is a consequence of the PLT process, that concrete nouns have relatively clear meaning schematics, that is, relatively stable inclusion/exclusion criteria for whether an object is a member of the word category or not. Thus, mistakes are rarely made in the communicative use of concrete nouns. Conversely, the absence of PLT (or triangulation) in abstract nouns, results in fuzzier meaning schematics and fluid inclusion/exclusion criteria. Consequently, abstract nouns can prove problematic to communicate and are subject to severe disagreement; an important point that is further developed with respect to the self and PLTb in Chapter 4.

Next, moving beyond isolated word meaning, several possible criticisms of similarity-based accounts were considered, most notably the compositionality requirement presented by Fodor. This requirement was shown to be lacking, for it failed to explain referential meaning. In lieu of this failure, it was argued that grammar is required for referential meaning, in that,

grammar is a necessary step in the process towards achieving full referential meaning. In the final section of this chapter, this grammatical conception of referential meaning was linked to an influential computational parsing theory of language, which explains several important aspects of linguistic complexity and will linked to neurophysiological data in Chapter 5. For the moment, the important point to take from this u-model discussion is that there is good evidence to accept that chunks of grammatical structure are associated and elicited by individual words, and that these chunks of structure (potentially) impress a referential type upon its constituents. This referential-impress process, if it occurs, would be blind to the type of noun or lexical information entering into the structures, for grammatical-structure building is only sensitive to grammatical features, not lexical content. This theory is consistent with the argument above that referential meaning is a grammatical phenomenon, not lexical. Both of these ideas are further developed in Chapter 3.

In Chapter 1, the literature on the abstract noun 'self' was reviewed across the disciplines of philosophy, psychology and neuroscience. It was concluded that a possible methodological problem exists with the way abstract concepts are approached. The exact nature of this problem was alluded to, but not stated in detail, this is the aim of the next chapter. However, to develop this methodological critique, it was first necessary to establish a working theory of how certain aspect of words – in particular, nouns – work and how they are used to refer to things. The current chapter has attempted to develop these points on the basis on sound theoretical and empirical evidence.

3 NOUN AND CLAUSAL REIFICATION AS LINGUISTIC ARTEFACTS

3.1 Chapter overview

In this chapter, it is argued that certain forms of theorising lead to methodological problems. In particular, to a specific type of ontological misclassification (here referred to as 'noun phrase (NP) reification') and to subtle, unintended changes in propositional meaning (here referred to as 'clausal reification'). NP reification is a specific type of Rylean category mistake that occurs when an abstract noun is falsely considered to be of the concrete noun category. Below, it is argued that this reification potentiality arises whenever an abstract noun enters a referential NP configuration. This occurs because of reasons relating to the referential hierarchy (discussed in Chapter 2 and extended in §3.3) and the Vosse and Kempen (2000) u-model (also discussed in Chapter 2, extended in §3.4), which together, predict that lexical content is blindly integrated into already-established referential structure. The limited role of lexical content in the formulation of reference type, allows this reification process to occur. Clausal reification is a process that occurs when an abstract noun, embedded in an NP (i.e. NP reification), occupies the subject position of a copular clause that is making an essentialist claim. This situation transpires when the metaphysically-inclined theoretician attempts to produce an essentialist proposition with a grammatical configuration of the copular predicational variety e.g. propositions of the form 'X is Y'; where X is an abstract noun phrase. These types of configurations are commonplace in metaphysical philosophy, psychology and neuroscience (and other disciplines), where establishing essentialist properties are the prize target of a theory. NP and clausal reification are here conceived as linguistic artefacts, as unintentional products of the way language functions. That these linguistic artefacts occur, explains the diversity of seemingly plausible but contradictory positions we see in the abstract concept literature, which was demonstrated in Chapter 1 with respect to the self.

In §3.7, these linguistic artefacts are positioned in the wider context of the cognitive bias literature. This literature, spearheaded by Kahneman (2011) and colleagues, argues that the human mind is prone to make systematic errors in certain aspects of cognition, which can be captured through scientific observation. Grammar, conceived in this sense, has yet to be considered as a possible locus of vulnerability to cognitive bias. The diagnostic conjectures of this chapter constitute an attempt to begin this conversation. The specific suggestion being that the metaphysically-inclined theoreticians, across disciplines, in their desire to make essentialist claims, are snared by a linguistically orientated trap, a cognitive bias that subverts the meanings of their propositions, making their propositions and accompanying theories unintentionally quasi-tautological.

3.2 Category mistakes

A category mistake, according to Ryle (1949/2009, pp. 15-18), occurs when things of one ontological kind are judged and presented as if they belong to another ontological kind. One example of this might be if someone said: 'The pain in my toe is pentagonal'. This is a category mistake, because 'pain in my toe' is not the type of thing that can be pentagonal. The subject: 'pain in my toe' is of the wrong ontological kind to have a pentagonal shape predicated upon it. However, one qualification to this point, this sentence may only be accused of a category mistake if it was intended to be interpreted literally. This sentence could potentially function as a figurative speech act, but would require previously established contextual support for it to be a successful communication. The focus of this chapter and thesis more generally, is on propositions that are intended to be interpreted literally.

Ryle formulated the idea of a category mistake to undermine Cartesian Dualism, the position that the mind and the body are separate, ontologically existing, things. For Ryle, Descartes was making a category mistake when he postulated the existence of a mind, as something that has real existence beyond that which is manifest in behaviour: '[its] one big mistake of a special kind. It is, namely, a category mistake' (1949/2009, pp. 5-6). Ryle's famous hypothetical example of a category mistake asks us to consider some rather naïve individual being shown around a university campus. Paraphrasing this example, we are asked to imagine an individual that is shown the various academic departments, the library, the lecture halls, the museums, the gymnasium, the administration offices, as well as meeting staff and students, but at the end of the tour, the individual asks the following: 'You've shown me the various academic departments, the library, the lecture halls, the museums, the gymnasium, the administration offices and I've even met some staff and students, but where exactly is the university?'. Here, it is Ryle's contention that this naïve individual has committed a category mistake, because s/he has inappropriately taken the word 'university' – as part of the NP 'the university' – to have the ontology of a physical object above and beyond the individual aspects of a university (those explored on the tour). In other words, the university is taken to be a physical object, named by a concrete noun, whereas, it is an abstract concept named by an abstract noun. This example involves a specific form of ontological category mistake called 'reification', where an abstract thing is mistakenly taken to be a concrete thing.

Category mistakes, in the form of reification, are instrumental to the argumentative thread of this chapter, but Ryle barely develops his explanation of category mistakes beyond this hypothetical example. The finer details of how and why these mistakes occur are left

unanswered (Magidor, 2013, p. 10). This chapter elaborates on these details, arguing that reification (specifically, the NP and clausal types) arise because of linguistically-orientated, cognitive biases that arise unintentionally in the development of theoretical work, especially in essentialist, metaphysical theorising. More generally, these biases are artefacts of the way we use language and the way that language functions (developed below). When theoreticians ignore these aspects, unintentional consequences result. In this sense, Wittgenstein was correct when he said 'philosophical problems arise when language goes on holiday' (1953/2009, §38). However, to make a persuasive argument for NP and clausal reification, it is first necessary to reconsider and extend some of the points made in §2.4 and 2.5, on referentiality and grammar.

3.3 Hierarchical reference via grammar

In §2.4 and 2.5, it was argued that referential meaning – in which a speaker picks out something in the world and refers to it – occurs in the relations between words, never in individual words, for no individual word can refer in and of itself. The main grammatical configuration involved in referential meaning is the noun phrase (NP). The structure of an NP is standardly depicted as: [$_{NP}$ [D] [NP]]; with D referring to a determiner (occupying the [D] position) and NP to noun phrase (headed by the noun N), both of which can be optional or obligatory depending on the configuration and other grammatical factors at the sentence level. However, there are alternative conceptions of NP structure, such as that represented in the DP hypothesis (see Abney, 1987), where a complex nominal of the above kind would be considered a determiner phrase (DP), but these structural disagreements make little difference to the force of the unfolding argument, therefore, the former shall be assumed.

As argued in §2.4, when words enter into these NP configurations we are able to engage in referentiality (Hinzen, 2014; Hinzen & Sheehan, 2013; Longobardi, 1994, 2005; T. Martin & Hinzen, 2014). For example, the lexical content associated with the term 'cat' does not refer to a specific cat or some cats I saw, because an isolated word (such as 'cat') only captures the general conceptual/lexical meaning, it is only in the grammatical relation between words where reference becomes possible (e.g. 'a cat', 'the cat', 'this/that cat'). However, to be clear, the grammatical NP configuration only represents one step in achieving full referential meaning, arguably serving as the referential blueprint of the phrase. To perform a fully-fledged referential speech act, typically requires that this grammatical NP configuration be saturated with appropriate lexical content (discussed below) and be embedded in an appropriate clause or sentence, which itself needs to be part of a wider context in which all aspects of grammar,

including the world/people orientated aspects of pragmatics, are appropriate for the reference. All subsequent discussions of referentiality and NP configurations assume this multifactorial conception of referential meaning as a given, but if reiteration of this point is required, it will be simply referred to as 'appropriate context'.

Following on from the above, the forms of reference generated in speech closely depend on the type of grammatical configuration used (evidenced below). Specifically, grammar co-determines whether a reference is *generic/abstract/mass, indefinite, quantificational, definite, rigid, deictic or personal*. These forms of reference differ in referential strength, increasing from left to right in the above-italicised list. The generic/abstract/mass and indefinite forms are referentially weak in the sense that no particular thing is referenced, whereas the quantificational, definite, rigid, deictic and personal forms are strong in the sense that particular things are referenced. Moreover, as referential strength increases from left to right, so too does the level of grammatical complexity, such that there is a positive correlation between grammatical complexity and referential strength, which can be described in the form of a hierarchical structure (also evidenced below). Here, the use of the phrase: '[increasing] grammatical complexity', indicates a decreased dependence on the interior aspects (i.e. lexical aspects) of the NP in the performance of a referential speech act and a correlating increased dependence on the edge aspects of the NP (i.e. grammatical aspects). This edge/interior distinction (formulised as [Edge[Interior]]) relates to the following research: (Longobardi, 1994, 2005; T. Martin & Hinzen, 2014; Sheehan & Hinzen, 2011).

There are, however, possible confounding factors that can affect this approximate definition of increasing grammatical complexity (due to the multifactorial nature of meaning), one of which is discussed below with respect to proper names. Nevertheless, what this broad generalisation states is that the more the referential act depends on the edge than on the interior (i.e. the lexical aspects) the more referentially specific it becomes. So, as we move from a reliance on lexical content to a reliance on grammatical content, we see a corresponding shift in referential strength, from weak to strong.

Let us now fully demonstrate this point through an examination of the different forms of grammatical configurations (i.e. those italicised in the previous but one paragraph), from the referentially weakest to the referentially strongest. The example NP configurations are embedded in sentences and are highlighted in **bold**:

Before starting, it should be noted that referential NPs are not confined to the subject position of sentence, where the subject position refers to the NP occurring immediately prior to the verb of the sentence (in normal English word order) and is what the sentence is fundamentally about. However, although referential NPs are not confined to subject position, once a sentence is complete, the subject NP takes on a special status as the main referential expression of the sentence. For example, in the sentence 'John likes his sister', 'John' is the subject of the sentence, since it occurs immediately prior to the main verb of the sentence (i.e. 'likes'). Whereas, everything else in the sentence (i.e. 'likes his sister') is the predicate of the subject 'John'. The predicate of a sentence tells us something about the subject NP. In this example, the predicate is grammatically complex, embedding a referential NP (i.e. 'his sister), but this is not the subject NP. This embedded NP specifies the grammatical object, where an object is the entity acted upon by the subject (technically the Patient or Theme), in this instance, the person who John likes (i.e. 'his sister'). In short, the subject NP of a sentence plays a special role as the thing that the sentence is about; all other NPs present in the sentence are subordinate to the subject NP.

When an NP embedded under a copular verb is grammatically predicative, the NP elicits a property reading, failing to be referential (T. Martin & Hinzen, 2014, p. 96). For example, consider the following:

John is **a teacher** John is [$_{NP}$ [**a**] [**teacher**]]. *Predicative NP*

Here, in the predicative 'a teacher', no particular thing is denoted, nor is there a mass or generic reading, so the referential hierarchy does not even get started.

The weakest forms of referentiality begin with abstract/generic/mass references, which in English are characteristically signified by the lack of overt determiner (i.e. where the D of the NP is absent (∅); abbreviation used in proper name discussion, below). In this instance, denotation is determined by the interior aspect of the NP, the descriptive/lexical content in the N, for example:

(1) I expect **justice** I expect [$_{NP}$ [**justice**]]. *Abstract NP*

(2) He hates **snakes** He hates [$_{NP}$ [**snakes**]]. *Generic NP*

(3) She had **snake** for lunch She had [$_{NP}$ [**snake**]] for lunch. *Mass NP*

These abstract/generic/mass readings are indefinite, since no individualised thing or things are denoted. The type of substances invoked in these configurations are uncountable (e.g. *one sand, *two air, *three justices), where: * indicates ungrammaticality). In short, the absence of an overt determiner in an NP configuration gives rise to the weakest forms of referentiality.

At the next level of the hierarchy, where a determiner is added to an NP, certain forms of indefinite-specific reference become possible, that is, if the accompanying noun is a count noun (i.e. nouns that can be counted as a differentiated single unit and can, thus, occur in singular and plural form). For example, with the addition of the weak determiner 'a' in the D position we are able to descriptively refer to a specific instance of a thing, but in an indefinite manner (Hinzen, 2014, p. 238), for example:

(4) **A snake** is chasing me [$_{NP}$ [**A**] [**snake**]] is chasing me. *Indefinite NP*

The use of the term 'weak determiner' above relates to research produced by Milsark (1977). Milsark argued that weak determiners, such as 'a', figure easily in predicative positions, whereas other determiners (i.e. 'strong determiners', such as 'the') do not:

(5) There is **a salesperson** at the door.

(6) ? There is **the salesperson** at the door.

Where '?' denotes that the sentence is infelicitous rather than outright ungrammatical.

With the use of 'a', a specific snake can be referenced, but not identified; though indefinite NPs do not have to be specific. Moreover, all forms of reference involving an empty or weak determiner would be translated in logical semantics with an existential quantifier, hence they reflect quantificational interpretations, such as:

(7) I saw **a snake** I saw [$_{NP}$ [**a**] [**snake**]].

Ex (x: snake & saw (I,x)), where 'E' is the existential quantifier

On the next rung of the hierarchy, after abstract/generic/mass and indefinite, we have definite references, which are typically marked by the selection of the strong determiner 'the' in the D position. In this instance, specificity is the default, assuming appropriate context. In the following example, a specific snake is intended, which would need to be familiar in the speech context, or have been mentioned previously:

(8) **The snake** is hungry [$_{NP}$ [**The**] [**snake**]] is hungry. *Definite NP*

Following on from the definite references, one level up the hierarchy, we have rigid designators/proper names, for which there is no determiner in English. However, despite lacking a grammatical function word, cross-linguistic evidence from (Longobardi, 1994, 2005; T. Martin & Hinzen, 2014; Sheehan & Hinzen, 2011), suggests that in this case, the lexical core of the DP is vacated, and head-movement by substitution (N-D) takes place, so that the N occupies the D position:

(9) **Victoria** is buying a snake

[$_{NP}$ [[**Victoria**] [$N_{\emptyset}$]] is buying a snake. *Rigid NP*

Here, in (9), 'Victoria' shifts from N to D position. In line with the remarks above, it is not always the case that proper names move into the D position, but if they do not, and the D position is filled, a descriptive or predicative reading is derived:

(10) No **Victoria** should be ashamed of her name

[$_{NP}$ [[No] [**Victoria**]] should be ashamed of her name

Here, the proper name 'Victoria' is not referring to a specific individual, but to set of individuals called 'Victoria'.

That proper names place high in this grammatical hierarchy, as grammatically complex phrases, may seem like a relatively provocative proposition, but, as stated above, the inclusion of proper names at this level of the hierarchy relates to the N-D movement as an instance of grammatical complexity (i.e. an increased dependence on the edge aspects of the

NP). To be clear, the reason that rigid reference is postulated to be grammatically more complex is that the proper name shifts into the D position. This type of movement is here considered to be an instance of increased grammatical complexity, because the core NP structure is adjusted; the lexical part shifts position from the interior to the grammatical edge, where the grammatical function words normally appear ('some', 'the', 'a' etc.).

There are two important criticisms of this view: the first questions its conception of grammatical complexity. On this point, it is worth reiterating the empirically observable phenomena discussed above, that in abstract/generic/mass/indefinite readings, grammatical function words (determiners) must or can be absent, hence this is a less grammatical form of reference, more lexically rooted. While in definite forms the grammatical function words are required (i.e. an increased reliance on grammatical aspects). In fact, in the case of deictic and pronominal forms of reference (detailed below), the lexical part is NOT required anymore and the phrase relies wholly on the grammatical aspects at this level of referential meaning. Finally, and most importantly, the specific details and levels of hierarchy explicated here need not be absolutely correct for the purposes of the unfolding argument, it need only be the case that reference is constrained by grammar (i.e. the particular, rule-governed relations between words), which abides by some form of hierarchical structuring, from weak to strong. That these necessities hold is difficult to deny.

The second possible criticism is more technical. Specifically, a critic may agree with this conception of grammatical complexity, but might argue that if N-D movement is an indication of grammatical complexity, then why doesn't this effect standard nouns as it does proper names, such as 'Snake' in the phrase 'Snake is on the menu'? In response to this, it is argued (T. Martin & Hinzen, 2014; Sheehan & Hinzen, 2011) that the N-D movement occurs in this example as well, and that this process has increased the referential strength of the configuration in much the same way as the N-D mechanism does for the proper name. An argument that the N-D movement has occurred is that a determiner cannot be slotted in front of 'Snake' in '**Snake** is on the menu' without fundamentally changing the referential form of the NP (e.g. '**A snake** is on the menu'). This impossibility is immediately explained if N already occupies the D position, hence no determiner can be added.

The reason proper names and common nouns are not normally expected to come with equal referential strength, post N-D movement, is because of confounding factors elsewhere in the multifactorial construction of referential meaning in language. Specifically, in this instance, the factor relating to frequency expectations. Proper names are expected to function

referentially and so are less expected to function descriptively. Whereas, the reverse is true of nouns like 'snake', in which the expectation is that they are more likely to function predicatively or descriptively, requiring that the speaker override this expectation, to generate the kind-referential effect that we see in the above example (e.g. 'Snake is on the menu'). Expectancies aside, the same grammatical mechanisms can apply to either make a proper name predicative (that is expected to be referential), or a common noun referential (that is expected to be predicative).

The systematic importance of this point is that, since all proper names can be used predicatively (i.e. when a determiner occupies the D position) and all nouns can be used kind-referentially (when they move into D), then plainly predicativity and referentiality are not primarily lexical properties of the noun, but spring from the relations between words, the grammatical configurations in which the nouns appear (i.e. in NPs) and in the wider grammatical context in which these NPs are embedded (i.e. either in subject, object or predicate position). In consideration of this, we may conclude that referential meaning is only partially modulated by lexical content, one of multiple contributing factors, including the referential/structural aspect under discussion here, set by grammar.

Returning to the hierarchy, and moving up to the next level, we see the introduction of the deictic element into in the D position. Of the various deictic elements, the demonstratives: 'this' and 'that' rank next in the referential hierarchy. These terms require contextual support by necessity: note that 'this' effectively means 'the + here', while 'that' means 'the + there', thus there is a greater complexity in demonstrative NPs over definite NPs, due to the principle of additivity. The contextual support required for deictic NPs is typically satisfied by some form of gesturing towards the thing being referenced, time locked to the deictic item (i.e. occurring in sync with the use of the deictic determiner). With the introduction of these elements, we get the ability to perform a maximally definite form of reference; perfectly coinciding with a weakening of the need for lexical information (i.e. demonstratives do not require a noun complement, see (13)). When deictic forms of reference occur, they should remove the need to ask follow-up questions such as 'which one?' examples include:

(11) **This snake** is called Jake [$_{NP}$ [**This**] [**snake**]] is called Jake. *Demonstrative NP*

(12) **That banana** is brown [$_{NP}$ [**That**] [**banana**]] is brown. *Demonstrative NP*

(13) **That** is a banana [$_{NP}$ [**That**] [N$_{\emptyset}$]] is a banana. *Demonstrative NP*

Of the various deictic items that exist in English, some are more grammatically complex than others. For example, 3rd person pronouns, like 'he' or 'she', lexically express Gender, but unlike the above-mentioned deictic/demonstratives, forbid an N-complement (e.g. * 'He man'), examples include:

(14) **He** is here [NP [**He**]] is here. *3rd person NP*

(15) **She** is away [NP [**She**]] is away. *3rd person NP*

2nd and 1st personal pronouns (such as 'you' and 'I', respectively) are devoid of Gender, forbid an N-complement and are positively specified for grammatical person, which no other NPs are. The linguistic phenotypes of 1st, 2nd and 3rd pronouns show a reduction in the necessity of lexical information, representing strong cases of reference. In these forms of reference, the link between grammar and lexical information is severed; partially so in the case of 3rd person pronouns due to some residual morphological/lexical information relating to Gender, and completely so in the 2nd and 1st pronouns. Therefore, due to their highly grammatical nature, the 2nd and 1st person pronouns occupy the top two positions of the referential hierarchy, with the 1st person pronouns on top.

As indicated above, 2nd and 1st person pronouns are predominantly (but not exclusively) indicated by the words 'you' and 'I', respectively, and are NPs in that they occupy largely the same grammatical positions as lexical NPs, as in:

(16) **You** are dangerous [NP [**You**]] are dangerous. *2nd person NP*

(17) **I** am a man [NP [**I**]] am a man. *1st person NP*

Both types of reference are considered deictic, because they require contextual support by necessity. However, in this instance, and in contrast to the other deictic terms, the mere utterance of 'you' and 'I' is often sufficient to establish the referential meaning, especially with the use of 'I'. So, unlike the demonstratives, there is no obligatory need for gesturing, but nevertheless these can be useful tools in disambiguating referential meanings. Note, however, that disambiguation is rarely (if ever) required when an individual uses the NP: 'I', but is

required by default with the 2nd person pronoun 'you', when there are two or more individuals involved in the discourse.

Being able to refer to oneself using the 1st person pronoun 'I', in English, is a unique and irreplaceable aspect of referential language, which facilitates a range of unique sentence meanings and aspects of cognition. A great deal of theoretical positions in philosophy, psychology and cognitive neuroscience lean heavily on this feature of language, especially those interested in the self. We will return to this in Chapter 4 (especially §4.3), but for now, we shall resume the thread of this chapter, starting with a recap of what has been established so far.

There are two main, methodological ideas under development in this chapter, termed: 'NP reification' and 'clausal reification'. On the former, the intention is to go beyond the remarks made by Ryle and explain how certain ontological category mistakes occur at the lexical and NP level. On the latter, the intention is to demonstrate how certain metaphysical propositions, those that make essentialist claims using a copular clause (i.e. of the form X is Y), with an abstract NP in subject position, are quasi-tautological. The argumentation so far has served to collate the necessary assumptions required in order that these ideas can be fully explicated. Specifically, it has been argued that:

1. Referential meaning arises in the relations between words (e.g. in NP grammatical configurations) and not in individual words.
2. That the type of grammatical configuration that occurs regulates the strength of the referential expression.
3. That the more grammatically complex an NP, the stronger the form of reference.
4. And finally, that these referential forms can be captured by a discrete hierarchical structure, from weak to strong (i.e. *abstract/generic, indefinite, quantificational, definite, rigid, deictic or personal*).

Next, these configurations are linked back to the Vosse and Kempen (2000) model of parsing, as a representation how this grammatical/referential configurations may be produced and comprehended psychologically. Importantly, the referential hierarchy, when considered together with the u-model, predicts that the construction of a referential expression is blind to

lexical content that enters it. Writing on this connection and the facet of blindness leads directly into the two main ideas discussed above, namely, NP and clausal reification.

3.4 The u-model and its relation to NP configurations

In §2.5, it was proposed that the Vosse and Kempen's (2000) u-model may accurately explain the mechanisms and computations responsible for the construction of referential expressions in speech production and comprehension. To recap, their u-model suggests that chunks of grammatical structure are stored in the brain, associated with, and activated by, individual words (i.e. the u-model holds that grammatical structure is stored in individual words and is elicited as a word is spoken/comprehended). The explicit suggestion in the present chapter is that these chunks of stored structural frames may correlate with the differing forms of grammatical configurations and referential meanings discussed above, typically activated by the grammatical function words in the D position, the edge aspect of the [edge[interior]] formal conception of the NP phrase (Sheehan & Hinzen, 2011). In the context of comprehension, the situation would go as follows: as words unfold one by one in a sentence, they elicit structural frames in the listener's u-space, which in turn constrain the types of words and grammatical structures that can follow. This assembly of parse-tree structure begins with the first word of the sentence, builds throughout each word of the sentence, and ends with the last word. For a sentence to be grammatical, there are preconditions in structural assembly that need to occur. Specifically, the individual grammatical structures, elicited word by word, must be grammatically compatible with one another. This involves the checking of grammatical agreement features, such as: number, gender, tense, case and person, as well as word order checks. Theoretically speaking, the binding of these grammatical structures only requires consideration of these grammatical aspects, in that the computational process of structure building does not care much about anything other than grammatical information. Importantly, this detail implies a disregard of, or disinterest in, lexical content or noun meaning. The grammatical mechanism involved in the construction of an NP and accompanying referential form barely interacts with the lexical noun content that enters the configuration, if one is present. The focus in the remainder of this chapter is mostly on NP configurations where a lexical noun is present in the N position, and a grammatical function word is present in the D position.

Further support for the lack of interaction between the grammatical mechanism and the lexical content is demonstrated by the fact that we can formulate grammatically correct

sentences that are semantically (or lexically) nonsensical, the famous example being Chomsky's sentence: 'Colourless green ideas sleep furiously' (1957/2002). This point is consistent with the empirical literature on semantic illusions (Barton & Sanford, 1993; Erickson & Mattson, 1981; Sanford, 2002), which posits shallow dipping as the standard approach to dealing with lexical information, discussed in §2.3. Moreover, modern neuroscientific research demonstrates a neuronal distinction between syntactic and lexical/semantic processing; for recent material on the matter see: Bastiaansen and Hagoort (2015) and Lam, Schoffelen, Uddén, Hultén, and Hagoort (2016), more on this in Chapter 5.

In respect of the above points, grammar can be said to function in much the same way as a mathematical system. In mathematics, a variable (say, 'x') is operated on in accordance with some mathematical rule or principle (e.g. $x^2 = x \cdot x$). In mathematics, the content of x is unimportant to the computational processing of these rules, in that x could represent any number and this would not affect the mathematical procedure, for the mathematical mechanism does not interact with this type of information, beyond some general constraints (i.e. the content of x is constrained to a number). In like manner, the lexical content (found in an NP, via the N) is irrelevant to the grammatical and referential computation of: $[_{NP}$ [D] [NP]], to the referential blueprint of a phrase. The implication here is that the NP computational/referential procedure and the processing of lexical content associated with an N, function in isolation from one another. This proposition can only be said to be largely true, not completely true, for mass nouns are quite restricted in their distribution, in that they do not function well with certain determiners (i.e. * 'a sand'). So, there is some shallow dipping involved to establish count and mass distinctions; however, one may want to argue that these are grammatical aspects, not lexical. Nevertheless, the critical point about these isolated processes relates to the grammatical indifference to lexical content, which allows abstract nouns, such as 'mind' or 'self', to enter into an NP practically unvetted. This is possible because, some abstract nouns have taken on count properties, so are able to slot in configurations better designed for concrete nouns (this idea is fully developed in §3.5 and 3.6, below).

Expanding on the consequences of the above, in conjunction with the u-model, the idea is that whenever an NP occurs in a sentence, it is the first item in the NP (the function word in D position) that determines the referential form that the NP will take, because the structure of the NP is stored and elicited by the edge, which sets the overall referential blueprint for the phrase. Following on from this, it is possible that through repeated uses of NPs in speech, specific D position items (e.g. 'a', 'the') become associated with specific referential

forms and are stored in memory as such; stored as part of Vosse and Kempen's (2000) chunks of structure. Essentially, this is equivalent to saying that the D position item or edge content, in the [$_{NP}$ [D] [NP]] structure, primes the referential form that will unfold, in much the same way as other priming effects function (Meyer & Schvaneveldt, 1971; Stanovich & West, 1983). Importantly, the referential blueprint for the phrase is set prior to the interaction with lexical information, which enters the N position.

As an example of this idea, imagine you comprehend the indefinite article 'a' in grammatical subject position of a sentence. In this situation, this will typically ('typically', in a statistical sense) signify the unfolding of an indefinite referential meaning. Here, it is possible that an indefinite referential structure will be elicited, which the succeeding noun will enter, irrespective of the type or stability of the lexical content it possesses; as long as it accords grammatically with the determiner. In like manner, 'the' will prime a definite referential meaning; 'this' a deictic referential meaning; 'He' a third person referential meaning; and 'I' a first-person referential meaning, but note that in the latter two cases, the structure is both primed and completed with the onset of the individual item. Furthermore, and equal to this, the absence of a grammatical function word, where a common noun occupies the D position, likely indicates the unfolding of a kind reference, as in the reading of 'man' in the sentence: 'man is dangerous'.

Analogous to a mathematical computation, the NP grammatical computation operates in isolation from the lexical content that enters into it. Once the structure is in place (i.e. when the D position is comprehended and the referential structure is set), it does not matter what subsequent content enters into the frame, as long as it does not violate grammatical agreement or word category features. What this means is that once the NP structure is determined in the D position, theoretically any noun can enter into the N position as long as it has agreeable grammatical features. In this scenario, the lexical content plays a limited role in the formulation of referential meaning. The referential form is first dictated by the structure, indicated by the grammatical information elicited in D position, and the lexical content, upon entry into the NP configuration, merely serves to push or pull the direction of this referential form towards a certain thing associated with the descriptive elements of the lexical content, assuming appropriate context.

The ability of lexical content to pull or push a referential structure in a certain direction, in a successful manner between individuals, depends on the stability of the lexical content. The more stable the lexical content, the clearer the boundaries and the easier it is to

apply the lexical schematic to a thing in the manner dictated by the referential form, which can easily be interpreted by speakers. If the lexical content is fuzzy and vague, then the noun fails to push or pull the referential form in any direction. In this sense, stable lexical content in the [NP [D] [NP]] structure is the fuel required for the referential form to connect with a thing or things. Conversely, unstable and vague lexical content represents the absence of fuel and leads to the inability of a referential form to connect with a thing or things. By analogy, a car cannot function correctly without fuel, likewise a referential form cannot function correctly without stable lexical content. The fuel-less car and referential form are left to idle in one place, metaphorically speaking.

To reiterate this key point, the lexical content of the noun – though a vital instrumental factor in achieving referential meaning (which is multifactorial) – cannot affect the form of reference that will occur, the lexical content merely manoeuvres the direction of referential form from within the boundaries of the referential structure set by grammar. The type of reference (i.e. the referential blueprint) is established through grammar, but if lexical content is present and stable, then this dictates the direction of the reference to a particular thing or things. If lexical content is present but unstable and vague, then reference can be said to malfunction. Of course, lexical content is not necessary for reference to occur, as in the use of pronouns, but as stated above, these are not of central interest at the moment; [NP [D] [NP]] structures are.

At this point, it is important to consider the consequences of these isolated mechanisms and facts about referencing, by examining what happens when different types of nouns enter into these configurations. If you do not like surprises, the answer is 'possible reification', when the noun is abstract.

Of the two types of nouns, let us first consider the concrete ones. Concrete nouns, as part of referential NP expressions, are ubiquitous in ordinary language, present in almost every sentence and essential for effective communication. As explicated in §2.2, the meanings of concrete nouns are derived from the physical world (i.e. argued to be acquired and maintained through PLT). Consequently, their meanings have a reasonable level of stability across speakers of a language. Alternatively, one might also say that their meanings have a reasonable level of stability across all speakers participating in a particular language game, where 'language game' refers to the Wittgensteinian concept that meanings in language are established through use, as a social activity between members (1953/2009). Hereafter, I will

use the term: ‘speaker(s)’ to denote individuals that are relatively synchronised in their concrete noun meanings, and are able to produce and comprehend them in the typical, intended fashion.

Stability of meaning is important in this context, because it allows referential meaning to function successfully; that is, a referential meaning, constructed and intentionally expressed by a speaker, which is accurately comprehended by others. For example, if I ask an office colleague to pass me **a stapler** (indefinite specific reference type) and they proceed to pass me **a stapler** then my referential expression has functioned correctly. However, if the colleague, on my request for **a stapler**, instead passes me **a banana**, then my referential expression has failed. In everyday communication, it is unlikely that this type of failure would occur, because both ‘stapler’ and ‘banana’, due to the PLT process, have largely distinct meanings that are stable across speakers. Therefore, an individual’s application of these schematics should easily differentiate between these two objects; unless some bright spark invents (or has already invented) a stapler that looks like a banana. Failure to differentiate between concrete nouns, in the above sense, does occur from time to time, when there is significant overlap in meaning schematics, or some other contextual feature allows for ambiguity. In this instance, speakers will provide modifications until the correct referential meaning is understood (e.g. ‘Not a banana, I wanted a stapler.’). In short, to make a successful reference to a thing – where the referential blueprint is set by grammar – the thing must have a stable and approximately uniform meaning across speakers, in order that the reference can be interpreted as intended. This is frequently achieved in the use of concrete nouns, because of the PLT process, but what of their abstract counterparts?

In §2.2, it was also argued that abstract nouns have unstable meanings that often vary across speakers. This is because abstract nouns, by definition, are lacking the physical properties that would facilitate PLT, the stabilising force behind noun meaning; however, there are other stabilising forces available to abstract nouns, discussed in Chapter 4. To reiterate, in order for referential meaning to function successfully in communication, the speaker’s referential meaning must be interpreted as intended. Given the nature of abstract noun meaning, one cannot be certain that this process has occurred without significant clarificatory discussion between speakers. For example, imagine I ask an office colleague whether they think **justice** was served in some controversial court case. Suppose they answer ‘yes’ to this question, and that this response mirrors my opinion on the matter. In this case, as it stands, there is no way to check whether my referential use of ‘justice’, that was active in my mind during the utterance of the question – which itself is likely vague and fluid – is the same conception (meaning

schematic) of justice that is elicited in my office colleague's mind during comprehension, which again, is likely vague and fluid.

Whenever we use an abstract noun, then, the content is transferred in a vague and fluid way, because these are inherently vague and fluid things (almost completely so), due to lack of PLT. When others comprehend our uses of abstract nouns, and we theirs, there is no instantly-available feedback to confirm that they have been comprehended as intended, further iterative clarifications are always required, but typically not asked for. Furthermore, on immediate reflection on our own referential use of abstract nouns, it quickly becomes obvious that we do not know what we intend to refer to when we use an abstract noun. However, as discussed above, argumentation in Chapter 4 attempts to reverse this scepticism towards abstract noun use by introducing a different means of stabilising their content.

3.5 NP reification

Earlier in this section, it was argued that when a noun enters into a referential NP configuration, it takes on the referential form (the referential strength), dictated by the grammatical signifier present in the D position. Furthermore, it was argued that the grammatical NP computation barely concerns itself with lexical/noun content, instead functioning in abstraction, much like a mathematical system. From this, it follows that when a lexical noun enters into grammatical configurations, its lexical content enters practically unvetted. Therefore, it is possible that lexical content might enter NP's that is ill fitted to the referential form dictated by the structure. Not concrete nouns, though, since these work well at any level of a standard [$_{NP}$ [D] [NP]] structure, assuming appropriate context:

(18) A man/cat/bus/pub — always indefinite.
(19) The man/cat/bus/pub — typically definite.
(20) This man/cat/bus/pub — typically deictic.

Tentatively, one might hypothesise that these structures are optimally designed for concrete noun entry in the N position; the use of 'designed' was not intended to convey an essentialist proposition nor to say that language had a designer, but was merely used for lack of a more appropriate term. Any concrete noun can enter in the above configurations, and will uniformly bring about approximately the same type of reference, depending on the configuration, in a

completely grammatical fashion. The same cannot be said of abstract nouns, for these do not effortlessly fit into these configurations, as the following examples show:

(21)	*a justice, *an intelligence	unresolvable as a reference
(22)	the justice, the intelligence	resolvable with complex descriptive context
(23)	this justice, this intelligence	resolvable with complex descriptive context

On (22), a resolution to the reference could be established with the following type of complex descriptive content:

(24) **The justice** administered by the British government was appropriate.

This linguistic behaviour is because most abstract nouns function more like mass nouns than count nouns. In fact, it is only the abstract nouns that most resemble count nouns that fit neatly into the [$_{NP}$ [D] [NP]] structure. In this instance, these abstract nouns perform similarly to their concrete counterparts, taking the (literal) referential impetus dictated by the grammar, yet without the above-mentioned ungrammaticality or infelicitous reading found with the mass-like abstract nouns. For example, 'a self', 'a mind', 'the self', 'the mind', 'this self', 'this mind' are all perfectly acceptable [$_{NP}$ [D] [NP]] configurations. In other words, for some reason (or reasons) certain abstract nouns develop the feature of being countable and are able to function in [$_{NP}$ [D] [NP]] configurations. It is at this point, where reification at the lexical NP level occurs, because the abstract noun, which is inherently vague and fluid in meaning, is able to take on the referential blueprint dictated by definite [$_{NP}$ [D] [NP]] constructions. Although a construction of this type (i.e. [$_{NP}$ [D] [NP]] with an abstract noun embedded in the NP) will sound grammatical, the abstract noun does not have the necessary stability of lexical meaning/content to drive the referential form towards an individuated thing agreed upon by language users, though one use may vaguely resonate with groups of users working in the same paradigm.

Relating this back to Ryle, recall that reification is a type of category mistake that occurs when an abstract thing is mistakenly treated as if it were a concrete thing. Based on the argumentation above, we can now speculatively infer that reification at the NP level occurs when certain abstract nouns take on a count noun interpretation. When this occurs, these

abstract nouns are able to slot into referential configurations [NP [D] [NP]] (e.g. indefinite, definite and deictic ones), which all require a clearly and stably individuated lexical schematic in order to function correctly as a referential expression.

The count-normalisation process that occurs in certain abstract nouns is likely the by-product of a certain pattern of language use, where speakers and theorists characteristically use the term with a count interpretation, as is the case with the majority of theories of self and many other concepts; including the often-related concepts 'the mind' and 'the soul'. Specifically, speakers and theorists, in their conversations and theories of self, implicitly or explicitly postulate a conception of self as a discrete countable thing. Moreover, this pattern of use implies the existence of multiple selves. In English, we also use the term 'self' as an adjective modifier of individuated things (e.g. self-refer, self-deprecate). All of these factors further cement the idea that the thing the noun refers to is discrete and thus countable.

All of that said, the prevalence of this linguistic phenomenon (i.e. instances of Rylian-type naïve individuals reifying at the NP level) is so low that it can hardly be considered problematic. In ordinary language, we certainly use these abstract nouns in referential configurations from time to time, but predominantly they are used in a vague and fluid sense, in which our intended referential meaning is either saturated by prior or subsequent sentential context or is intentionally (or unintentionally) produced and comprehended as a relatively unimportant aspect of a sentence and thus is skipped-over in conversation. Note, however, that both approaches (i.e. further saturated or simply skipped over) would lead to miscommunication if the language produced was scrutinised in detail. Furthermore, in ordinary language, our typical intended use is not one in which a strong ontological, metaphysical, essentialist claim is being constructed and communicated. In fact, as will be argued hereafter, it is exactly when these philosophical intentions are present, where the truly problematic aspects of abstract nouns and reification surface. Specifically, when the above-explicated abstract NPs are embedded in essentialist propositions in the grammatical subject position. The focus of this chapter is when this occurs in copular clauses/propositions, hereafter discussed under the name: 'clausal reification'.

3.6 Clausal reification

A clause is the smallest grammatical configuration that can function as a sentence and that can express a proposition (i.e. truth and falsity). All clauses and propositions possess a subject and predicate. As discussed in §3.3, the subject is what (or whom) the sentence is about, whereas

the predicate tells us something about the subject, thus a predicate is an expression that can be true or false in its relation to the subject. In English, a copular clause is a subject-predicate grammatical configuration connected via one of the following copular verbs: 'is', 'am', 'be', 'are', 'was' and 'were'; arguably, these terms are different inflections of the copular verb 'be', but this debate will not be pursued since it does not alter the import of the current argument. The principle function of these verbs is to link the subject and predicate of the clause. For the purposes of the argument in this section, the focus is on copular clauses functioning as propositions, linked by the copular verb 'is'; but note that the arguments made below are equally applicable to the other copular verbs. Hereafter, example copular clauses, copular sentences or copular propositions (terms used interchangeably) will be indented, with the subject NPs highlighted in **bold,** predicates in braces { } and the copular verbs (as part of the predicate) highlighted in *italics*. As an example of the above notation, consider the following copular clause:

(25) **The boy** {*is* tall}

Here 'The boy' is the subject of the sentence, the thing which the sentence is about, whereas {*is* tall} is the predicate, which is telling us something about the subject. '*is'* is the copular verb heading the predicate, which links the two aspects of the sentence (i.e. the subject and predicate). Considered together, we (potentially) learn something about the boy, that he is tall.

The subject of a copular clause is usually an NP, whereas the predicative part of a copular clause can be a prepositional phrase (PP), an adjective phrase (AP) or another NP. For example:

(26) **The snake** {*is* under the table} *PP*
(27) **The snake** {*is* dangerous} *AP*
(28) **The snake** {*is* my pet} *NP*

According to Mikkelsen (2011) – based on the foundational research performed by Higgins (1979) – there are four types of copular clauses in English: 'Predicational', 'Specificational', 'Indentificational' and 'Equative'.

Predicational copular clauses predicate a property onto the thing that the subject refers to, as in (25), (26) and (27).

Specificational copular clauses specify who or what someone or something is, rather than to say something about the thing. The subject introduces some descriptive content and the predicate provides the reference for the descriptive content. An approximate parallel to this can be found in mathematics, such as in the expression: 'x = 7', where: 'x' is an arbitrary symbol/variable that is not fully specified and thus can take on content through instances of it, '=' functions like the copular verb, and '7' functions like a referential thing.

When 'x' is assigned the value '7', 'x' becomes a bound variable, where a bound variable refers to a variable that was free but is now bound to specific content. In ordinary language, specificational clauses look as follows:

(29) **The director of Pulp Fiction** {*is* Quentin Tarantino}
(30) **The murderer** {*is* that man}

Equative copular clauses equate the referents that flank either side of the copular verb. For example:

(31) **Hesperus** {*is* Phosphorus}
(32) **Peter Parker** {*is* Spiderman}

'Hesperus' and 'Phosphorus' are two different, yet referentially equivalent, names for the celestial planet, Venus. Whereas 'Peter Parker' and 'Spiderman' name the same fictional person, but in this instance, the different names have somewhat different lexical information associated with them.

Finally, identificational copular clauses are characterised as possessing a demonstrative pronoun or phrase in the subject position, which must be understood to have a deictic reference, thus requires contextual support, such as gesturing. For example:

(33) **That snake** {*is* a Mamba}

(34) **This man** {*is* David}

According to Higgins (1979), the subject of identificational clauses (i.e. the demonstrative pronouns or phrases) are referential, whilst the predicate is identificational. These clauses are typically used to teach people the names of things. For current purposes, identificational copular clauses are hereafter conceived as a subset of predicational copular clauses.

For the present argument, the important types of copular propositions are the predicational and specificational ones; therefore, the focus will narrow onto these forms. One way to delineate whether a copular clause is predicational or specificational is to examine the strength of the referential content across the copular proposition, for predicational and specificational clauses have contrasting phenotypic asymmetries between the strengths of the referential content in the phrases that flank either side of the copular verb. Specifically, a predicational copular phrase is referentially stronger in the subject position in comparison to the post-copular position, which often functions purely descriptively. Conversely, a specificational copular phrase is referentially stronger in the post-copular position compared to the subject position. Consider the following examples (also further detailed in Table 1, below):

(35) **John** {*is* the mayor} *predicational*

(36) **The mayor** {*is* John} *specificational*

In (35), 'John' is referentially stronger than 'the mayor'; assuming that in the context of this utterance, it is clear who 'John' is, across speakers. This is the case, because 'John' refers to a real individual known across those speakers involved in the utterance, whereas the use of 'Mayor' in this context, simply evokes lexical content and is not directly referential to a person or thing in the world, certainly not to the extent that 'John' is. Referential flank asymmetries of this type are predicational:

John {*is* the mayor}

John > the major

Where '>' equals: 'greater than' in the referential scale

In contrast to this, (36) has 'The major' occupying the subject position and 'John' occupying the post-copular position, thus eliciting the reversal of (35), in terms of asymmetrical referential strength. Referential flank asymmetries of this type are specificational:

The mayor {*is* John}.

The mayor < John.

Where '<' equals: 'less than' in the referential scale

	Pre-copular	**Post-copular**
Predicational	Referentially stronger: John	Referentially weaker: the major
Specificational	Referentially weaker: The mayor	Referentially stronger: John

Table 1: Demarcating between predicational and specificational copular clauses, based on referential strength in the pre- and post-copular copular phrases.

When the referential strength of the two flanking NPs is equal we get an equative copular clause:

(37) **Hesperus** {*is* Phosphorus}

Hesperus = Phosphorus

Where '=' equals: approximate symmetry in the referential scale.

Here, 'Hesperus' and 'Phosphorus' are two different, yet referentially equivalent, names for the celestial planet, Venus.

Why are these copular clauses and their associated details important? The answer is that these sentential structures represent the grammatically simplest way to construct propositions that can make essentialist/metaphysical claims. As such, these structures are frequently (but not exclusively) used in theoretical prose that attempts to explicate the fundamental nature of things. For example, imagine a deeply thinking theorist (immersed in

the essentialist/metaphysical method and mind-set) stating that some concept, let us call it X, is essentially constituted of A, B & C, such that:

X {*is* A, B & C}

Here, A, B & C need to function predicatively, stipulating some essential and necessary properties of X, because this is what the theorist intends to tell us about X, that its essential nature is comprised of the properties A, B & C.

To express this point in ordinary language, using a copular grammatical construction with a literal intention, requires that the construction be of the predicational type. For, it is only in the predicational copular clause that the predicate is telling us something informative about the subject, the application of which can be true or false in an essentialist sense. However, it need not be essentialist. In some cases of predication, that we will here call 'naïve', there is no essentialist implication, as exampled in (38):

(38) **The man** {*is* walking}

Here, we have a predicative copular construction, which is truth apt, but not in the essentialist sense, typical of theoretical assertions. Naïve copular clauses merely predicate a property onto a subject in a transient sense, where the predicate is true or false in application, for a temporal slice of the subject's existence. In order for a copular clause to have a non-transient essentialist quality, it is necessary that the predicate be of the 'individual-level' or 'kind-level' predicational type, as defined by Carlson (1977, 1980) and not a 'stage-level' predicate, where the predicate is true or false at a particular temporal stage. Hereafter, 'individual-level' shall be referred to as: 'transient' and 'stage-level/kind-level' as 'intransient'.

Let us now reconsider the implications of the above argumentation on the differing forms of nouns, in the context of reification, with a special focus on the theories of self outlined in Chapter 1, summarised in Table 2, below:

Theory Number	Theorist	Summary Proposition
S_1	Plato	The soul/self is immortal and immaterial.
S_2	Aristotle	The soul/self is a form or essence of the thing.
S_3	Augustine	The soul/self is immortal and immaterial, but intimately connected with the body.
S_4	Descartes	The mind/self is immaterial, interacting with the body through an aspect of the brain.
S_5	Locke	The self is psychological continuity (e.g. consciousness and memory).
S_6	Hume	The self is a bundle of perceptions.
S_7	Siderits	The self does not exist and a person is a conceptual fiction that is useful, up until a certain point.
S_8	Kant	The self is not an object of experience, but is a transcendental condition for it.
S_9	Gergen	The self is a social construct, residing in the communicative relation between individuals.
S_{10}	Dennett	The self is a narrative fiction.
S_{11}	James	The self has four aspects; material, social, spiritual and pure ego.
S_{12}	Neisser	The self has five aspects; ecological, interpersonal, extended, private and conceptual.
S_{13}	Gallagher	The self is a sense of agency and a sense of ownership functioning correctly, in conjunction with a narrative.
S_{14}	Churchland	The self is a folk psychological concept, which will ultimately be displaced and explain through neuroscience.
S_{15}	Northoff	The self is an individual's relation to exteroceptive stimuli in the world, mediated through self-referential processing. This processing is a functionally independent module in the cortical midline structures of the brain.

S_{16}	Metzinger	The self does not exist, it is an illusion brought about as a consequence of certain processes in the brain, However, a self schematic does exist.

Table 2**:** Summary of self positions examined in Chapter 1.

At the grammatical level, the self positions in Table 2 all adhere to the predicational copular form, with definite NPs in the subject position and a mixture of complex and intransient NPs and APs in the predicate position. This should not be especially surprising since $S_{1\text{-}16}$ were rendered as such in Chapter 1. Nevertheless, it would be difficult for the respective authors or their modern followers to deny that these propositions hold to their work. Moreover, it would be difficult to make or summarise these positions in a grammatical construction other than a copular clause, without losing their precision and essentialist qualities.

As argued in §1.6, these individual theories all make relative sense when considered in isolation but are largely contradictory when considered as a set. We now have the relevant details in place to understand why this is so. Specifically, propositions $S_{1\text{-}16}$ are all making essentialist, metaphysical claims about what the self is, via the predicative copular clause form. However, for a copular clause to be truly predicative in meaning, it is required that the left flank, the pre-copular phrase, be referentially stronger than the right flank, post-copular phrase, as described above. *Prima facie*, this appears to be the case for $S_{1\text{-}16}$, since formally speaking these configurations are grammatically predicative. However, due to the nature of the abstract noun 'self', its vagueness and fluidity, this predicative grammatical form is subverted in these propositions, such that the opposite interpretation is dominant (i.e. the specificational type). In other words, the propositions $S_{1\text{-}16}$, though predicational in grammatical form, are undermined elsewhere in the referential process, subverted by one of the other factors that contributes to referential meaning, namely, the lexical content, which bears on the referential import of the subject NP. More specifically, the lexical content in the right flank is more exact and stable than that present in the left flank. In this sense, these propositions behave more like specificational copular clauses, than predicational.

Let us isolate one of the above examples from Table 2 – namely, Hume's (1739/1888) theory of self – to demonstrate this point in detail. The proposition associated with Hume:

S^6 = The self is a bundle of perceptions

To recap, Hume rejected the notion of a perfect self, persisting through time. Instead, arguing that what we think is the self (responding to a Cartesian conception) is merely a consecutive stream of perceptions that are inconceivably rapid and transient in nature (i.e. a bundle of perceptions). S_6, above, is a summarising proposition that captures this perspective, and is rendered in the familiar grammatical form below, in (39):

(39) **The self** {*is* a bundle of perceptions}

At the grammatical level, (39) is a copular clause of the predicational variety. We have a definite NP grammatical structure in the subject, pre-copular position, which is initialised by the grammatical function word 'The', as part of the theoretical extension of the Vosse and Kempen model (2000) described in §3.4. Furthermore, we have a predicate, in the post-copular position, that is grammatically indefinite and intransient in nature, initialised by the grammatical function word 'a', thus satisfying the copular predicative and essentialist requirements of a metaphysical proposition and is consistent with taxonomy of copular sentences explicated by Higgins (1979) and Mikkelsen (2011).

Grammatically speaking, then, (39) is making a predicative, essentialist, truth-apt claim in a very unproblematic fashion. The problem with this configuration, though, is not at the grammatical level of the proposition, but in the lexical aspects that saturate the grammatically dictated referential form or blueprint. Recall that, in §3.4, it was argued that lexical content (amongst other things) drives the referential form (set by grammar) to a thing or things in the world. The NP 'The self' of (39), has the grammatical features that typically give rise to a definite reference, but possesses the lexical content that is vague and fluid for reasons described in §2.2 and 3.4. In this instance, the definite NP collapses into an underspecified, referentially weak phrase, functioning somewhat like an empty vessel, similar to 'x' in 'x = 7'. In contrast to this, the content on the right, post-copular flank of the copular clause (39) (i.e. 'is a bundle of perceptions'), though grammatically speaking is in predicative position, primed by the indefinite article 'a', has lexical content that referentially outweighs that present in the subject NP 'The self'. Importantly, the use of 'outweighs' here refers to increased stabilised content (i.e. the meaning schematic underpinning the phrase is more sensitive than the subject NP). Specifically, the lexical content in the post-copular clause – that

which is derived from the terms 'bundle' and 'perceptions' – is subject to PLT (see §2.2), thus has a level of meaning stability beyond that present in the abstract noun 'self'. So (39), which starts out grammatically as a copular predication, ends up, at the lexical stage of referential saturation, functioning more like a specificational copular clause, due to the relatively referentially-heavier right flank.

If this argument is correct, it has significant and equivalent consequences for the other metaphysical/essentialist approaches outlined in Table 2, as well as other essentialist propositions of this form found in other disciplines (psychology, cognitive neuroscience, etc.). Namely, that although appearing grammatically and propositionally predicative, they are behaving (propositionally) more like specificational copular clauses. As discussed above, specificational copular clauses specify who or what someone or something is, with the left flank providing some descriptive content and the right flank providing the reference for the left flank content; mathematically equivalent to 'x = 7'. If this is occurring in propositions $S_{1\text{-}16}$ then they can be said to be behaving quasi-tautologically due to the osmotic process that is occurring within the copular clause; 'osmotic' in a restricted (non-biological) sense of movement and merging of meaning within the clause. Fundamentally, as the processing of the copular clauses $S_{1\text{-}16}$ unfolds, the lexical content of the left flank, the pre-copular phrase (e.g. the subject: 'The self') is forced to align/merge itself with the content of the right flank, post-copular phrase (i.e. the predicates of $S_{1\text{-}16}$); forced due to the right flanks, stronger referential content and (relatively) more stable lexical content.

This osmotic merging of meaning is not a tautology in the traditional sense of the term, where you see a repetition of concepts either side of the copular verb (i.e. the same concept repeated using different phrases) as in (40) below:

(40) **Bachelors** {*are* unmarried men}

Instead, what is occurring in propositions $S_{1\text{-}16}$ is a quasi-tautological process ('quasi' here meaning partly/almost or similar to tautological), in which there is an asymmetrical osmotic merging of meaning. Relating this back to (39) and further developed in (41) below, this osmotic merging of meaning can be described as a two-staged, spanning two factors of the multi-factorial process required for achieving full referential meaning; but note that this description and the example below are not intended to suggest a specific, linearly-organised

unfolding of referential meaning in a temporal sense. On (41), in the first stage (S1) of referential meaning, the grammatical structure suggests a predicative reading, due to the referential forms within the structure, as outlined in previous paragraphs. Whereas, in the second stage (S2) of referential meaning, the asymmetry in lexical meaning stability suggests a specificational reading. The interaction (INT) of factors S1 and S2 results in an osmotic merging of meaning, where X aligns itself with the value of Y, resulting (RES) in a quasi-tautological state of Y = Y:

(41) **The self** {*is* a bundle of perceptions}

S1:	X is Y	Predicative copular clause at the grammatical level
S2:	x is **Y**	Asymmetry in lexical meaning stability
		(i.e. right flank is heavier/stabler than left flank)
INT:	X ← Y = Y	Osmotic merging of meaning, X displaced by Y
RES:	Y = Y	

Classical tautologies and the above outlined quasi-tautologies are not fallacious in and of themselves. They can be used in ordinary language for emphasis and pragmatic/conversational convenience, as in (40). However, they do become fallacious when they are mistakenly taken to possess non-arbitrary truth-values, such as those present in essentialist, metaphysical propositions that aim to capture true and false aspects of something in an intransient sense. The contention of this chapter is that exactly this process is occurring in the positions outlined in Table 2, as well as many other theoretical positions, across disciplines, which follow this form.

However, the problem of abstract nouns goes beyond the limited form outlined above in $S_{1\text{-}16}$, in which the abstract noun is part of definite grammatical NP, accompanied by a determiner in D position. On closer examination, this osmotic merging can be said to extend to all copular clauses where an abstract noun occupies the subject position, irrespective of whether a determiner is present or not. For example, in (42), (43) and (44) below, the subject NP of the copular clauses have no determiner in D position; though arguably the N is occupying the edge position in these [Edge[Interior]] structures, since it unlikely that sentence-initial positions have empty categories. The predicates: 'X', 'Y' and 'Z' should be interpreted as symbolising

complex and intransient NPs or APs in the post-copular position, which give rise to a heavier right flank, relative to the vague and fluid subject NP, in the left flank:

(42) **Justice** {*is* X}
(43) **Beauty** {*is* Y}
(44) **Knowledge** {*is* Z}

If sentences of these types are taken to possess non-arbitrary truth values, in an essentialist, metaphysical, intransient sense, then these sentences can be said to be equally subject to the osmotic processes and consequences specified above, as described in (41). Specifically, these propositions, though grammatically predicative, are functioning specificationally, due to the factor of lexical content; the result, quasi-tautological propositions in the fallacious sense. In other terms, (42), (43) and (44) is equivalent to stating: 'Of all the things that could be Justice/Beauty/Knowledge, Justice/Beauty/Knowledge is… X/Y/Z' This particular representation highlights the weaker quantificational/indefinite, variable-like, flavour of the left-flanking NP, while the right-flanking phase provides the instance.

To sum up this point on reification, it has been argued above (and in Chapter 2) that referential meaning arises in the relation between words, being present in a range of NP configurations, which systematically fit within a particular level of a hierarchal structure, dictated by the complexity of the grammar involved. These NP configurations are the first step towards achieving full referential meaning; they constitute and set the referential blueprint of a referential utterance. The grammatical mechanism underpinning these NPs, function in relative isolation from the lexical content that enters them. On a closer examination of this mechanism, it is clear that concrete nouns (i.e. nouns that derive their lexical information from things in the world, established through triangulation) function successfully in definite NP configurations and are the foundational basis for effective communication. In addition to concrete nouns, some abstract nouns can also figure into definite NP configurations, even though they are not derived from things in the world and therefore have vague and unstable meaning across and within speakers. The result of this possibility is reification at the NP level, captured by the naïve individual example and their misuse of the NP 'the university'.

The prevalence of NP reification, in the above restricted sense, is negligible and thus can hardly be considered problematic. However, at the clausal level, reification in abstract

nouns becomes more prevalent and problematic, especially in theoretical prose where essentialist, truth apt, literal propositions are intended. The particular clause examined in this chapter was the copular clause with an abstract NP in subject position, connected to the predicate by the copular verb 'is'. When these types of clauses are used with the aforementioned essentialist, propositional intention, a kind of osmotic reification occurs that subverts the meaning of the clause, from the intended and required predicational meaning (X = Y) to the unintended and unwanted specificational meaning, thus making these propositions and the theories built upon them quasi-tautological (Y = Y).

An important follow-up question to the above conclusion is 'why are we (humans) able to make mistakes of this type?' To which, one possible response might be to argue that the language faculty is not optimally 'designed' to carve up reality in a rational, logical and essentialist manner as a set of simple primitives that compose into complex facts; that which is typically required for clean metaphysical theorising. Instead, the language faculty has (perhaps) evolved for reasons more geared towards communication, as defended by the later Wittgenstein (1953/2009), Austin (1962) Searle (1969) and more recently by Scott-Phillips (2014), amongst others. Consequently, language may not have the optimal structure for metaphysical theorising. On this view, language is a tool; say, a hammer. A hammer is useful for hammering a nail. When a hammer is used in a non-hammering task (e.g. as a precise instrument in an invasive surgical procedure) it is likely that errors may occur. Likewise, language appears to have evolved to have a particular use, possibly as a basic communication system, and when language is used beyond this use (i.e. in precise metaphysical reasoning or conceptual analysis, which requires several assumptions criticised above and in Chapter 2), then here too errors may occur, such as NP and clausal reification. In this context, the language system can be said to have a set of biases infused within it, much like other aspects of human cognition. A lack of attention to these biases lead to linguistic artefacts in theoretical prose, such as those described above. In the final section of this chapter, these biases are positioned in the wider context of the now well-established cognitive bias literature.

3.7 NP and clausal reification as cognitive biases

Since Tversky and Kahneman (1975), we have known that the human form of cognition is prone to make systematic errors in certain situations. Consequently, the idealised conception of the human mind, as the paragon of pure reason and logic, is incorrect. More recently, Kahneman (2011) has provided substantial arguments to support the view that there are two

systems at work in the human mind, called ‘System 1’ and ‘System 2’ and it is the structure and interplay of these two systems that gives rise to these systematic errors or cognitive biases.

System 1 operates rapidly and automatically with little to no conscious effort or agential control on the part of the individual. Analogously, System 1 functions much like the autonomic nervous system, but with a focus on regulating aspects of cognition, as opposed to the regulation of bodily functions. For example, System 1 regulates aspects of cognition such as: the orientation to and identification of sensory cues in the environment, the creation of associations in memory, the interpretation of emotions in faces, automated motor responses (e.g. during walking, running and riding a bike, etc.), in the performance of simple calculations (e.g. 2 + 2), making stereotypical judgments, and relevant for this chapter and thesis, the production and comprehension of grammatically simple sentences. Note also that the initial acquisition of language is a product of this System 1, given its automaticity, in the developing child.

In direct contrast to System 1, System 2 operates on complex computations and decisions that require a concerted attentional effort from the individual, as such, they are experienced as agentially-controlled processes (i.e. as concentrated efforts of our ‘free will’ capacity). System 2 is identified with aspects of cognition such as: considering the validity of an argument, focussing attention on a particular voice in a noisy room, performing complex calculations (e.g. 86 · 72), engaging in moderate/intense exercise, playing chess, counting the amount of words on a page, and so on. In all instances of System 2, the common denominator is attentional effort required from the individual. Consequently, in the relevant cognitive activities associated with System 2, you will perform less well, or not at all, if your attention is inappropriately directed, or you are in the grip of System 1 processing.

Kahneman (2011) provides significant empirical evidence to suggest that we humans are dominated by System 1 processing, even in cognitive activities where System 2 would be more appropriate, suggesting that this occurs because of a human psychological predisposition to find mental effort aversive. In other words, we are cognitively lazy and tend to rely heavily on System 1 processing. This point is particularly important for the current consideration of grammar as a possible locus of vulnerability to bias. Whilst Kahneman acknowledges that language – being effortlessly acquired, produced and comprehended – is part of System 1, he does not consider whether there are systematic errors in the use of certain grammatical structures, in fact, to my knowledge, the discussion of this chapter is the first to consider this as a possibility.

A grammatical mechanism that only shallow dips into lexical information can lead to systematic errors, such as those outlined in this chapter. One possible reason why we do not register these errors is because we rarely attend to our use of words and grammar in a way that would draw the necessary attention to them. Of course, in serious writing (academic and non-academic), much thought and time goes into how best to communicate a thought to a reader, on how best to phrase a sentence or proposition and how sentences and propositions relate to one another to make a coherent argument. Nevertheless, it is not clear that adequate consideration is given to how language functions, when conceived as a limited tool. Translating this into the Kahneman framework, the idea is that our System 1 use of language involves using language freely, as it comes, in an inattentive manner. Part of this System 1 use, requires that one hold to the assumption (implicitly or explicitly) that language is perfectly transparent, that it can be unproblematically applied to a variety of tasks, including the metaphysical, essentialist ones. Whereas, in reality, language conceived as a limited tool, potentially geared towards effective communication, requires a System 2 level of attentional energy be dedicated to it, especially when attempting to use language beyond ordinary communication (e.g. when it is used as a tool to carve the world up into metaphysical primitives). Allocating mindful attention on the nature of how language functions, whilst simultaneously developing theoretical ideas, will help circumvent and attenuate the effects of the above-outlined linguistic biases. In short, when the inattentive System 1 use of language is in play during theory construction, systematic linguistic biases such as NP and clausal reification can occur. To avoid this problem, theorists should consider engaging an attentive, System 2, perspective to language. Dedicating attentional effort to thinking about how language works helps in the avoidance of the inbuilt linguistic traps that have here being argued to permeate the human cognitive phenotype.

3.8 Chapter conclusion

The argument of this chapter began with an examination of category mistakes. Specifically, the category mistake in which an abstract thing is ontologically mistaken to be of the concrete category, a fallacy termed ‘reification’. After outlining how reference occurs in grammatical relations, it was then argued that reification occurs because the grammatical mechanism that constructs the referential blueprint of a phrase operates in relative isolation from the lexical material that enters into it. This is typically unproblematic when the noun is concrete, deriving its content from PLT, but becomes problematic when the noun is purely abstract, devoid of PLT-derived content. Next, it was argued that when these reified abstract nouns enter into

subject NPs of copular clause configurations, which are making essentialist truth claims, they subvert the propositional meaning of the clausal configuration from predicational to specificational. If this conjecture is correct, then a failure to acknowledge the osmotic alteration in meaning may result in the unknowing production of quasi-tautological propositions and theories. A quasi-tautological theory would likely be internally consistent, with respect to its explication of the concept of interest, but would also likely be contradictory with other explanations of the same concept, this discrepancy is exactly the state of affairs we see in the theories of self across philosophy, psychology and neuroscience, outlined in Chapter 1. Nevertheless, being quasi-tautological does not necessarily detract away from the importance or beauty of a proposition or theory, it merely highlights the fact that the proposition/theory is not making a valid, synthetic, essentialist, metaphysical claim.

The final aspect of this chapter tried to place the outlined linguistic artefacts into the wider context of cognitive biases to which we humans are subject. This placement is largely speculative and underdeveloped in its current form, for the position does not have the empirical support typically required for acceptance in the cognitive bias literature. It is hoped that subsequent work could build upon this idea.

In terms of traditional philosophy, this chapter (and the previous chapters) may read as anti-metaphysics, but it is not intended to be simply this. Instead, at a fundamental level, the intention is to motivate a reorientation in methodological approach to metaphysics, towards a different and arguably more stringent means of doing metaphysics, that takes seriously the scientific examination of language, as well as principles of psychology; and later, when sufficiently advanced, neuroscience. In Descartes' *Meditations* (1641/2008), he tells us to doubt everything that we do not hold to be certain and indubitable, but in doing so, forgets to question the very tool he used to express this point, namely, language. In this chapter, it has been argued that an examination of this tool is necessary if we are to avoid unconscious and unwanted biases. Based on the above argumentation and observations, language – the medium by which theoretical prose is created – should no longer be considered as transparent (i.e. made up simple essential primitives) and thus needs to become the centre of attention in disciplines beyond the sub-branch of philosophy that examines language. In other words, if one accepts the proposition that language is non-transparent, then one should prioritise the question of 'what is meaning?' over the questions 'what is the self', 'what is the mind?', 'what is knowledge?' and so on. To sidestep meaning at the outset will result in the production of biased theoretical prose, populated with unwanted linguistic artefacts.

Arguably, there is a correspondence between what is said in this chapter and what Wittgenstein states in his *Tractatus*, namely: 'What can be said at all can be said clearly; and whereof one cannot speak thereof one must be silent' (Wittgenstein, 1921 Preface) also: 'The right method of philosophy would be this. To say nothing except what can be said, *i.e.* the propositions of natural science, *i.e.* something that has nothing to do with philosophy…' Both of these propositions relate to Wittgenstein's picture theory of meaning, in which a proposition is said to have sense if its structure pictures reality ('pictures' in a broad sense) and is truth apt. Based on the argumentation of this chapter, it could be argued that propositions consisting of concrete nouns (in subject position of copular clauses) are able to picture reality more easily than purely abstract nouns, due to their possession of temporally extended physical properties that, through triangulation, give rise to increased meaning stability, relative to their abstract counterparts. For example, it is easier to picture and to determine the truth or falsity of the proposition: 'The cat is sat on the mat', which consists of two concrete NPs, versus the proposition: 'The self is consciousness' which consists of two abstract NPs. Based on this comparison, it could be suggested (speculatively) that there are two possible ways to proceed hereafter: either [1] we only state things that can be said clearly, specifically, propositions with concrete NPs embedded in them, for their truth or falsity are easier to determine, or [2] we should endeavour to stabilise abstract NPs so that they can be sensibly used in propositional constructions, with clear truth conditions. To my mind, option [1] is too defeatist. Moreover, there is much at stake with the analysis of abstract concepts (e.g. work on abstract concepts has significant implications on ethics and law making). Therefore, the next chapter will attempt to develop a positive theory of self, only with an adjusted methodological approach that accounts for the non-transparent aspects of language. This theory will hark back to a fleetingly mentioned idea of Chapter 2, called 'perceptuo-linguistic triangulation of behaviour' (PLTb).

4 WHAT NOW? THE 'SELF' STABILISED IN BEHAVIOUR

4.1 Chapter overview

In the previous chapters it was argued that abstract nouns need to be stabilised in order for them to be both useful and synchronised (in their use) across a linguistic community. Perceptuo-linguistic triangulation (PLT) is an inappropriate stabilising force, because abstract nouns, by default, do not have the standard physical object that would facilitate triangulation and thus stabilisation. However, it is argued below that perceptuo-linguistic triangulation of behaviour (PLTb) may be an appropriate alternative, for abstract nouns can be linked to behaviours, which have an observable, shared existence in the perceptual domain and thus are susceptible to triangulation and stability; here, the use of the term 'behaviours' is used in a non-standard sense, referring to a broad range of actions that are produced by individuals, organisms, objects and/or systems, which all have an observable shared existence in the human perceptual domain; fully explicated in §4.2.

One of the major difficulties in this PLTb approach is isolating the behaviours of interest (i.e. those indicative of selfhood) without resorting to a metaphysical, essentialist approach to concepts, argued against in Chapter 3. Using PLTb, in conjunction with basic scientific methods and certain foundational assumptions in statistics – namely, the Gaussian distribution and central limit theorem – it is possible to circumvent this trap, for this approach leans on and extends an established scientific approach.

The 1st person pronoun 'I' is considered as a possible behavioural indicator of self, but after critical consideration, it is found that 'I', used in isolation, is largely inadequate. In its place, fully socialised and statistically typical utterances are proposed as a core constituent behaviour in the profile of self; where 'utterances' are conceived as self-standing, syntactically independent, units of discourse, providing new information. Related, and in addition to this, it is then argued that evidence from neuroscience, on language comprehension, may also have a role to play in this profile of self, a point which primes the empirical investigations in Chapter 5. The neuroscientific discussion in this chapter involves a preliminary examination of the N400 and P600 components of an electroencephalography recording. These components reflect different cognitive aspects of the language comprehension process and can illuminate aspects of the self in their relation to language.

4.2 The triangulation of behaviour

In §2.2, it was argued that there are different ways in which the meaning of a word can be acquired and maintained. Specifically, concrete nouns (i.e. nouns that name everyday physical

objects) are acquired through PLT, where there is a three-fold, cross-modal, perceptual interaction between two or more individuals and an object in the world. In contrast to these, abstract nouns (i.e. nouns that typically name ideas, states or qualities), lacking in a clear third-point physical object, are acquired through point-2-point (P2P) interactions, in which information is exchanged between individuals on the noun in question. There are several problems associated with the use of the latter types of nouns, some of which were detailed in Chapters 2 and 3. However, as was briefly highlighted in §2.2, P2P information can, and often is, supplemented by a form of triangulation, in which the abstract noun is associated with a behaviour, a process referred to (in this thesis) as PLTb. In PLTb, the problems associated with abstract noun meaning stability, dissipate (somewhat), for the triangular aspect of PLTb is a stabilising force (further detailed below). The example given of PLTb was of a teacher explaining the abstract noun 'politeness' to a child (child1). The teacher, in an attempt to explain what politeness means, draws the attention of the child1 to the behaviour of another child (child2), where child2 is observed saying: 'Please' and 'Thank you' in some social interaction. Here, 'politeness' is predicated onto child2's behaviour (i.e. in the saying of 'Please' and 'Thank you'), and this behaviour, observed by child1 and the teacher, plays the equivalent role of the physical object in the PLT process.

The ability to associate an abstract noun with a behaviour, allows a triangular process to take place. In PLT, the triangular interaction is between two or more speakers and an object in the world, whilst in PLTb the interaction is between two or more speakers and a behaviour in the world. The triangular process is important, because it allows the meaning of a word (its lexical content) to stabilise so that it can be used between speakers successfully, that is, communicated and understood as intended. Stabilisation can occur because behaviour has a shared existence in the perceptual domain and thus, can be subjected to the triangular process. Frequent, consensus-driven triangulation of behaviour, results in a stabilisation of lexical content, for the constant stream of references to behaviour, uniformly associated with a particular word (between speakers), incrementally strengthens the sensitivity of our semantic/lexical schematics. The analogy developed in §2.3 was that of the electrical force, as postulated in physics. Specifically, in the same way that the electrical force keeps the atoms of an object together, including its boundaries, so too PLT and PLTb keeps the boundaries of a word meaning, clear and stable. PLTb is a weaker stabilising force in comparison to PLT, for the behaviours triangulated in PLTb are more dynamic and complex than the objects triangulated in PLT.

A parallel of the above methodological approach – that is, of linking abstractions to behaviour (in the broad sense of the term) – can be observed in the physical sciences, whenever the subject matter exists outside the human, cross-modal, perceptual range. Take visual perception in humans, which is limited to a narrow window of the electromagnetic spectrum, at a macro level. Human beings cannot directly observe things that operate outside of this spectrum and/or beyond the limited macro view of the world, including atomic and subatomic particles. So, to circumvent this limitation, scientists generate inferences on the nature of these things indirectly based on predicted, observable behaviours within the human perceptual range; with advances in technology continually widening this observable range. For example, to better understand subatomic particles and the early state of the universe, particle physicists accelerate protons, in the Large Hadron Collider, close to the speed of light and place them on a collision course with one another, which leads to the dispersion of smaller, subatomic particle tracks (behaviours), which are then detected (observed) by advanced semiconductors. The statistical analysis of this behavioural data generates insights into the subatomic realm, such as the recent discovery of the Higgs boson.

This methodological approach can also be observed with respect to the scientific investigation of black holes (and gravity), the physicist postulates the existence of a black hole as an abstractum, a former star now collapsed, but which still exerts a significant gravitational pull; likely equivalent in strength to that present prior to the star's collapse (see Hawking, 2001 for overview). Conceptualised as such, physicists cannot observe black holes directly, via the above-mentioned PLT process (in more scientific terms, through light detection). Instead, they infer the existence of a black hole based on the (behavioural) effect it has on nearby physical matter (i.e. the downstream computed consequences in the observable range). For example, as a black hole moves it draws matter towards it, including planets and stars, a process referred to as accretion. In this scenario, matter accelerates as it moves closer to the centre of the black hole (the singularity), and as it accelerates it emits light that radiates back into space, which can be observed and measured via satellites and telescopes. The abstract NPs 'black hole' and 'gravity' (and 'atoms' and 'subatomic particles') are talked about, and theorised on, in a coherent sense across individuals (scientists and others), because they have been linked to an approximately agreed set of observable and triangulable behaviours, those mentioned above. Similar to 'politeness', the content of the NPs 'black hole', 'gravity' etc… are stabilised through PLTb. One possible explanation, as to the evolution of this process in science, is that, in the first instance, a set of unexplained behavioural phenomena is observed, and the scientist,

in their desire to efficiently communicate and explain these unexplained phenomena to others, bestows upon it an abstract name or label, to which the behavioural information is anchored. However, on the above, a disanalogy should be noted between black holes, atoms and the abstract noun 'politeness', in that, the former two, though conceived as theoretical abstractions are often referred to as real, physical objects, whereas, 'politeness' does not refer to a physical object. This difference does not affect the argument under development in this chapter. The usefulness of this analogy lies in the fact that in science, theoretical abstracts are stabilised in relation to overt behaviours all of the time, which often leads to clear advances in knowledge. Therefore, the proposal here, further developed below, is that this method could hold true for abstract terms/concepts in philosophy too, such as 'the self', where consensus on the true nature of the concept appears difficult to achieve via the traditional, philosophical methods (as detailed in Chapter 1, 2 and 3). To clarify, it is not the suggestion that there should be a science of abstract concepts, but that aspects of the scientific method could assist in the explanation of these abstract topics.

At this point, it should be emphasised that the use of the term 'behaviour', the 'b' of 'PLTb', is not intended to be interpreted in the everyday sense of the term, that is, as behaviour exclusively associated with humans (the typical subject matter of psychology). Instead, 'behaviour' is used here to refer to a broad range of actions that are produced by individuals, organisms, objects and/or systems, which all have an observable shared existence in the (human) perceptual domain. In addition to this, it should be noted that PLTb theory does not share any of the ontological or epistemological commitments associated with the philosophical position known as behaviourism. Philosophical behaviourism, broadly speaking, is the view that psychological states should be analysed as behaviours. For example, a behaviourist typically demands empirical evidence of behaviour to explain states identified by mentalistic vocabulary, such as 'belief' or 'desire' and if there is no clear difference in the associated behaviours, between these states then, for the behaviourist, there is no difference between the states. At the radical end of the behaviourism spectrum, such as that introduced by Skinner (1957/2014), it is argued that mental terminology, where it cannot be eliminated, should be wholly replaced with behavioural explanation, meaning that 'behaviour can be described and explained without making ultimate reference to mental events or to internal psychological processes. The sources of behaviour are external (in the environment) not internal (in the mind, in the head)' (Graham, 2017). Both PLTb and behaviourism emphasise the importance of behaviour with respect to abstract nouns, but the reasoning underpinning this mutual interest

is different. PLTb is motivated by facts of language (discussed in Chapters 2 and 3) and by foundational aspects of scientific methodology (discussed in this chapter), using behaviour to aid the development of abstract notions; here, the self.

4.3 PLTb, the self and I

The first step of applying PLTb to the self requires that an explicitly agreed-upon behaviour or set of behaviours be established that are indicative of a self. These behaviours are not limited in scope, they can span the full range of capabilities, experiences and/or phenomena, but for the purposes of explicating the notion of self associated with our species, it makes sense (at least initially) to focus on behaviour that is uniquely human, that is, behaviour that does not clearly overlap with non-human animals or other entities, more generally. With this approach, the theory of self that arises will capture important aspects of what it means to be a human self. When this theory is stable across a linguistic community, it can then be factored into important philosophical debates that hinge on a strong notion of self (e.g. euthanasia, abortion, advanced directives and so on).

Outlining behaviour in this way might appear to presuppose an understanding of the notion in question (i.e. to know what a self-behaviour is, first requires knowledge of what the self is), this is true in some limited sense, for one must have some initial grasp of the term 'self' to get started with this new approach, but this need not constitute a metaphysical, essentialist theory/understanding of the self, with a full complement of necessary and sufficient conditions. This latter presupposition of self only arises in those still engaged in the essentialist, metaphysical mind-set. An alternative to this, as discussed above, is a basic scientific approach in establishing the relevant behaviours, which need not resort to essentialist philosophy for epistemological justification. In fact, the potential behaviours that can be linked to the self can be fixed and understood in a relatively mundane and everyday manner and will not likely generate a diversity of contradictory perspectives, as was seen with the more traditional philosophical approach, detailed in Chapter 1. For example, a top contender for an example self-behaviour in humans might be our ability to self-refer, given that the term 'self' modifies the verb 'refer' and that self-referring is one of the defining characteristics of human communication, which, in its linguistic form, is exclusive to our species. In English, this self-behaviour, is canonically (but not necessarily/exclusively) produced via the utterance of the pronoun 'I', as one highly prevalent instance of the grammatical 1st person system.

Besides the reason outlined above, the importance of the term 'I' is emphasised across many theories of self and other philosophical topics, more generally. For example, Perry (1979), in his critique of the traditional doctrine of propositions, states that 'I' is an essential indexical that we cannot do without, a view also supported by Evans (1981). The general idea here is that sentences containing the first person (e.g. 'I') cannot be replaced in all contexts by sentences not containing it, therefore the first person is essential and that such sentences are also essential to intentional explanations. On this, one might extend Perry's point by saying that the pronoun 'I' allows certain types of beliefs and intentional behaviours to arise, which no other referential phrase could generate.

Perry's famous example is of a hypothetical event in a supermarket: Whilst shopping in a supermarket, Perry spots a distinct trail of sugar on the floor. Being a good citizen, he decides to follow it, so that he can inform his fellow customer of their dire situation, the torn bag of sugar. Despite his persistent sugar-trail hunting, he is unable to find the customer. But then, suddenly, he comes to the realisation that he is the person with the torn bag of sugar and that he has been cyclically following his own trail. This realisation coincides with the thought: 'I was the shopper I was trying to catch… I am [the one] making a mess' (p. 3). The salient point here is that 'I' is essential to these thoughts, for it cannot be replaced with any other referential expression and still have the same meaning and effect on the individual's beliefs and behaviour. In other words, without 'I', in the above scenario, there is no other way that Perry could have had the realisation that he was the one responsible for the mess, a realisation that in turn prompts his intentional behaviour to stop following the sugar trail and fix his mess. For example, imagine replacing 'I' in 'I am [the one] making the mess' with the proper name of the individual stating 'I', namely, 'John Perry'. Here, if Perry had the thought 'John Perry is the one making the mess', it is not clear that this thought brings about the same sort of realisation and behavioural change that 'I' does, and so on for all other grammatically viable NPs that can fit the grammatical subject position of the sentence. The reason why 'I' is essential, whereas 'John Perry' and other 3rd person NPs are not, is because the alternatives to 'I' do not directly refer to the speaker as the 1st person, whereas 'I' does. Knowing the name 'John Perry' and the associated description (e.g. a man, a philosopher, wears a beard) is not equivalent to knowing that I am John Perry. In English, this self-recognition is facilitated by 'I' as part of the wider grammatical person system.

There are, of course, positions arguing against the essential nature of the indexical (Cappelen & Dever, 2013). In particular, they argue that essential indexicals, such as 'I', are

just a case of words in opaque environments and that many expressions operate with, and are sensitive to, opaque environments. Stated differently, that 'I' and 'John Perry', in the shopper example, are simply two co-referential expressions, without this being known, thus, there is nothing distinctive about indexicals. To demonstrate this point, following on from the Perry example, it could be imagined that one, whilst following the sugar trail, could suspect that Clark Kent is the shopper making the mess, without knowing that Clark Kent's alter ego is Superman. In this scenario, it is Cappelen and Dever's point that finding out that Clark Kent was Superman, would bring about similar psychological consequences as to the realisation brought about by the essential indexical, that 'I am [the one] making the mess'. Stated differently, knowledge from the realisation of the identity statement (i.e. Clark Kent is Superman) facilitates new beliefs and actions, comparable with the realisation brought about by the essential indexical 'I', thus, making the latter inessential. It is difficult to engage in this criticism without getting embroiled in the game of abstract definitions (e.g. the definition of an 'essential indexical'), the very method argued against in Chapter 3. Yet, it is necessary to defend the idea that 'I', as a form a self-reference, occupies an important a place in the philosophy of self, and self-theories more generally. To lessen this tension, it should be sufficient to note that 'I', as an exemplar of a 1st person reference, as part of a grammatical person system, encodes person information beyond that present in, and achievable via, lexical content words, such as 'Clark Kent' and 'Superman'. Thus, this indexical still maintains a unique function, even if one concedes it is not unique to the level specified by Perry and others.

Along similar lines, Kaplan (1989) discussed two basic principles for indexicals: that they are context dependent and that they are directly referential. Indexicals are context dependent in that the reference of the indexical depends on the context of the utterance. For example, the indexical 'I' of the sentence: 'I am tired' denotes a different reference depending on who states it. If I state: 'I am tired' then the 'I' refers to me, whereas if someone else states it, then it applies to them and so on for the different types of indexical and users/contexts. Kaplan's second principle, that indexicals are directly referential, holds that once the reference is determined, it becomes a fixed part of the propositional content. The indexical may come to denote different objects across differing contexts, but in the propositional evaluation of the utterance, only the object referred to in context is relevant.

Another perspective where 'I' takes centre stage is in the psychological examination of self in its minimal form (Gallagher, 2000; White, 2015), discussed in §1.4. The minimal theories of self focus on the phenomenological experience of self as subject, devoid of temporal

extension. This minimalistic modification emphasises the stripping away of the unessential features of self to reveal its primitive foundation, the self in its most basic form. This notion has been explored in several ways. Of particular relevance for the current argument is Gallagher's approach, in which he emphasizes the significance of the pronoun 'I', stating that the use of 'I' represents immediate access to the minimal self. Moreover, he states that certain uses of 'I' are immune to error, meaning that when a person uses 'I' to refer to themselves, they cannot be mistaken as to whom they are referring to.

This idea, of 'I' being immune to error (through misidentification), is a feature of language noted by Wittgenstein (1958/1984) and was further developed by Shoemaker (1984) and others. Wittgenstein distinguished between two uses of the 1st person pronoun in self reference, namely, 'as subject' and 'as object':

> 'There are two different cases in the use of the word "I" (or "my") which I might call "the use as object" and "the use as subject". Examples of the first kind of use are these: "My arm is broken", "I have grown six inches", "I have a bump on my forehead", "The wind blows my hair about". Examples of the second kind are: "I see so-and-so", "I hear so-and-so", "I try to lift my arm", "I think it will rain", "I have toothache". One can point to the difference between these two categories by saying: The cases of the first category involve the recognition of a particular person, and there is in these cases the possibility of an error, or as I should rather put it: The possibility of an error has been provided for… On the other hand, there is no question of recognizing a person when I say I have toothache. To ask "are you sure that it's you who have pains?" would be nonsensical.' (pp. 66-67)

Building on this point – in the context of self-knowledge, perception and memory – Shoemaker introduces the idea and phrase: 'immune to error through misidentification' (i.e. the immunity principle) to capture the two distinct uses that Wittgenstein outlines: where 'I' as subject is immune to error and 'I' as object is vulnerable to error (Pryor, 1999). Further extending the immunity principle, Gallagher (2000) and Campbell (1999) explore the idea of its possible malfunction in patients with a diagnosis of schizophrenia. For example, thought insertion, a prominent symptom of schizophrenia, involves the patient experiencing thoughts that they do not identify as their own, instead, believing that they are inserted into their minds from some

external source. Thus, the idea would be that patients with schizophrenia and thought insertion may not make the typical connection with thoughts, including the referential use of 'I'; as subject and object, and thus have a diminished form of minimal self.

The point of this short review of 'I' is to demonstrate the prominence and importance of 'I' in theories of self and philosophy more generally. In consideration of these facts, it does indeed seem possible that self-referencing would be the type of behaviour that many theorists would be receptive to as a viable behavioural indicator of self. In other words, it is possible that many theorists would agree that self-referring with 'I' is a good indication of selfhood. However, for reasons now detailed, holding to the view that 'I' is a viable behavioural indicator of selfhood is, in itself, unsatisfactory. Primarily because a human may use the term 'I', but still be accused of possessing a diminished form of self (on the ordinary use of the term), as alluded to above. For example, imagine an individual who has had a stroke, leading to a deficit in language production (e.g. Broca's aphasia). In this scenario it is imaginable, but perhaps not especially probable, that the individual (post-stroke) is only able to say the word 'I', similar to Broca's famous patient Lebornge who could only utter the word 'tan' (see Dronkers, Plaisant, Iba-Zizen, & Cabanis, 2007 for historical overview). Alternatively, consider a patient with a diagnosis of schizophrenia who utters a sentence such as 'I am Jesus', when they are likely not Jesus (Hinzen, Rossello, & McKenna, 2016). Finally, consider the pronoun reversal phenomena in autism, where the individual, typically a child, confuses the terms 'I' and 'you' (Kanner, 1943; Naigles et al., 2016). Considering these scenarios, it is unlikely that the mere use of 'I' would satisfy as a consensus behaviour across theoreticians as a valid indicator of intact selfhood, for something more is required. Two highly related modifications are hereafter presented, and when these modifications are taken into consideration, a more robust behavioural candidate of selfhood arises, one more likely to generate adequate consensual agreement.

The first modification goes as follows: For 'I' to be considered a self-behaviour (a valid self-reference), it must be stated in a way that is fully socialised or normatively correct. Here, 'fully socialised' and the synonym 'normatively correct' refer to an individual's ability to play the language games present in their society, especially those in their immediate environment (Wittgenstein, 1953/2009). In this sense, the use of 'I' is more than an isolated, context-free word, but is viewed as a normatively correct move in a publicly played activity; where 'normatively' is used to refer to normal behaviour but does not attribute a moral value to this behaviour (further discussed in §4.6). By analogy, using 'I' is a bit like moving a piece

in a game of chess. To be said to be engaged in the game of chess, one needs to abide by the conventional rules when making a chess move (a behaviour). Likewise, to be engaged in self-reference, one must (more-or-less) abide by the rules of the wider language game when self-referring. If the majority of users does not follow the rules, then the game breaks down. That is, if 'I' is used in a non-normative, haphazard fashion, where it is not consistently used to self-refer, then the self-reference aspect of 'I', as part of a language game, is undermined.

The second proposed modification – intimately related to the first, but from a more technical viewpoint – is the requirement of statistical typicality. To be statistically typical requires that the word (here 'I') be used in such a way that it can be considered normal in a numerical, statistical sense. The approach captures the same phenomena as the 'fully socialised' modification, but in a statistical fashion. Explicating the full details of the statistically normal use of 'I' is not needed here, for the point can be made at a higher level, through an examination of the statistical model of normality, see Figure 3, below:

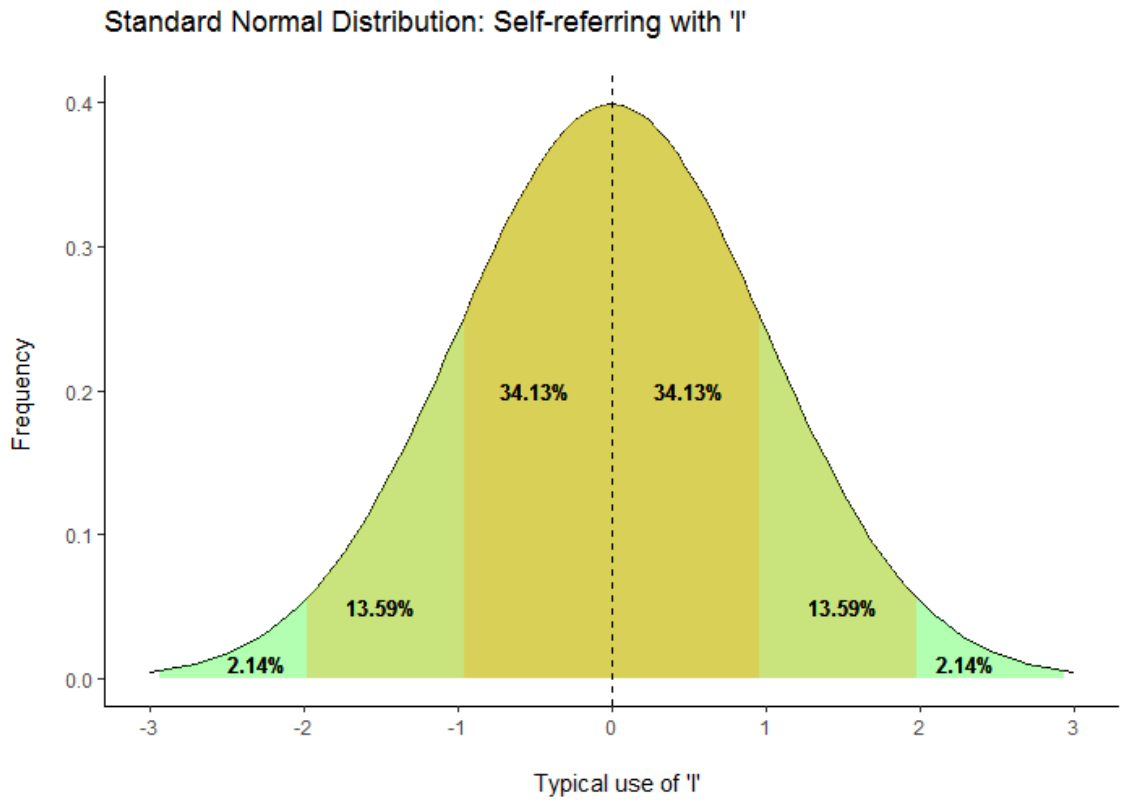

Figure 3: Simulated standard normal distribution curve to demonstrate the typical use of the pronoun 'I'.

Figure 3 is a standard normal distribution, also referred to as a Gaussian distribution or bell curve. The normal distribution is (arguably) the most valuable tool in inferential statistics, used

heavily in both natural and social sciences as a base reference for statistical comparisons. The reason the normal distribution is so important is that it provides an accurate model for observed frequency distributions for many events and phenomena that occur naturally in the universe. Take, for example, the measurement of weight in humans, which follows a normal distribution pattern similar to that represented in Figure 3. Most individuals will be approximately average weight, that is, within ±1 standard deviation (SD) of the mean (μ) (i.e. the two sections labelled 34.13%). In addition to this, there will be a set of individuals that are moderately lighter/heavier than the μ, represented between ±2 and ±1 SDs of the μ (i.e. the two sections labelled 13.59%). Finally, a relatively small number of individuals will be significantly lighter/heavier than the μ, represented on the curve as more than ±2 SDs from the μ (i.e. the two sections labelled 2.14%).

The pattern of normal distribution consistently arises in nature. Important to this pattern is the mathematical principle referred to as the Central Limit Theorem, which states that the sampling distribution of sample means is approximately normally distributed, irrespective of the distribution of the random sampling. Of course, there are exceptions to normal distributions that occur when the event in question is close to a boundary, such as floor or ceiling effects in some test or measure. For example, this might occur when many of the students in a test get close to 100%. In this scenario, the data will not be normally distributed but will be positively skewed.

Relating this model back to self-reference, it is probable that the use of the self-reference 'I' will follow a normal distribution pattern such as that depicted in Figure 3, with the majority (**95.44%**: 13.59% + 13.59% + 34.13% + 34.13%) of individuals falling within ±2 SDs of the μ, with a minority (≈**4.28%**: 2.14% + 2.14%) of individuals using 'I' in a non-statistical typical and non-normatively socialised manner; beyond ±2 SDs of the μ. Note that the threshold for typical and non-typical is approximately based on the standard significance value assumed in statistical tests (e.g. 5%). Furthermore, the normal distribution is asymptotic, whereby the curve approaches zero as the x-axis tends towards infinity, therefore, there is a certain percentage of the tails which is unaccounted for in the above description, but which is negligible for the purposes of the current argument.

Based on the above, 'I' can be considered as an adequate constituent behavioural indicator of self if it is being used in a fully socialised and statistically typical (95.44%) manner. The identification of a normatively correct and statistically typical use of 'I' – hereafter, jointly referred to as 'mid95%' (with reference to the middle 95 percentile) and the converse, a non-

normative, statistically non-typical use referred to as 'end5%' (with reference to the end 5 percentile) – could potentially be achieved via two methods. Firstly, some end5 use of 'I' will be strikingly obvious in ordinary conversation and will prove quite jarring to hear (e.g. the above-mentioned imaginary stroke patient who repetitively utters 'I'). Secondly, end5 use of 'I' will be observable and identifiable through the scientific examination of language, i.e. in linguistics. Linguistic analysis can take many forms, but in the context of this thesis, an important approach involves the comparison of speech across groups, in which the performance of linguistic analysis aims to identify patterns of language that can be used to differentiate between the groups and reveal features of cognition that may (or may not) be at play. For example, Fineberg et al. (2015) found that patients with a diagnosis of schizophrenia produced significantly more third person pronouns ('they') and significantly fewer first person singular pronouns ('I') in comparison to patients with a mood disorder diagnosis. There were substantial limitations associated with this particular study, not least their lack of a healthy control group and their questionable form of data collection; see Maatz (2014) for details. Yet this type of approach, the linguistic analysis of language output, is one means of distinguishing mid95% and end5% usage, but more sophisticated and ecological experiments would be required to demonstrate a clear distinction (see Çokal et al., 2018 as example); William Jones, the author of this thesis, is a co-author of this paper.

Unfortunately, there is still a looming criticism of this overall approach that needs to be addressed and adjusted for, namely, that 'I' can only be assessed for normativity and statistical typicality when understood as part of an utterance, which itself needs to be embedded in wider discourse. Why? Because words, as discussed above, are not used in isolation, in a vacuum, but are part of an intricately-played public game, where the minimum constituent piece of this game is not a word, but a contextually-appropriate utterance; where an 'utterance' is here defined as a self-standing, syntactically independent, unit of discourse providing new information; a criterion of which the referring term 'I', considered in isolation, does not satisfy. Normativity and statistical typicality are means of assessing whether a self-reference, embedded in an utterance, is a functional move in a language game.

It is only at the utterance level where we can judge the normativity and statistical typicality of a self-reference, whether its use is of the mid95% or end5% type. Based on the definition of utterance at play here, it can be said that all utterances have a referential element, but it is not the case that all utterances involve a self-reference e.g. 'Paris is the capital of France' is an utterance with referential elements, but none of which are self-referential.

Although, if I assert the above proposition, there would be an implicit self-referential element, namely, that I consider it true (more on this below). This idea, that not all sentences are self-referential, potentially reduces the number of utterances that we could link to the self and the usefulness of this behaviour as an indicator of selfhood. However, theoretical and empirical arguments can be made to demonstrate that self-reference occurs in all utterances, irrespective of whether an overt lexical pronoun is present or not. A further examination of this point will serve to clarify the nature of the PLTb solution to self. This examination will begin with an exploration into Tarski's work on truth.

4.4 Self-reference in utterances

In a period of relative scepticism towards the abstract noun/notion 'truth', Tarski (1944) embarked on a truth-restoration project, with a methodological focus on the application of the predicate 'is true' to a sentence. The theory that he developed could be categorised as a redundancy theory of truth, where, roughly speaking, redundancy theories of truth argue that stating that a proposition 'is true' is equivalent to simply stating the proposition; hence, why the 'is true' predicate is characterised as redundant. The precipitation of the redundancy perspective can be attributed to Aristotle (384-322BCE/1933, §4.1011b) in his well-known definition of truth: 'To say of what is that it is not, or of what is not that it is, is false, while to say of what is that it is, and of what is not that it is not, is true'. Later, Frege (1918/1977) and Ramsey (1927) further explicated this type of view.

In the historical context of Tarski's work, theories of truth typically fell into one of three categories: coherence theories, pragmatic theories and correspondence theories (McGinn, 2016, p. 148). Coherence theories hold that a propositional belief is true if it coheres with other beliefs in the individual's system; pragmatic theories hold that a proposition is true if it increases utility in the individual; and correspondence theories hold that a proposition is true if it corresponds to facts in reality. Of these positions, in the context of modern philosophy (and the modern world, more generally), correspondence theory is typically seen as the most robust, since it allows a role for the world in the construction of a true proposition, where coherence and pragmatic theories do not achieve this is a clear manner. However, correspondence theories of truth are not free from criticism, for it is often said of these views that they presuppose or equivocate the notion in question, for example, shifting the analysis of truth to the analysis of 'correspondence' or 'facts in reality'. Tarski's theory of truth was an attempt advance correspondence theory, whilst avoiding these problems.

In the first step of his argument, Tarski outlined two constraints that he thought were necessary for a theory of truth to be successful, namely, that it should be 'formally correct' and that it must possess 'material adequacy'. Formal correctness is the requirement that the theory be free from logical errors and should not presuppose, or equivocate on, the notion to be defined (i.e. truth), as discussed above. This constraint was a prerequisite for Tarski and many other theorists of the time, as part of the tradition looking for a formal, logical solution to meaning. The second constraint, material adequacy, is the requirement that a theory must capture the meaning of truth as it is used traditionally and ordinarily:

> 'Consider the sentence "snow is white." We ask the question under what conditions this sentence is true or false. It seems clear that if we base ourselves on the classical conception of truth, we shall say that the sentence is true if snow is white, and that it is false if snow is not white.' (Tarski, 1944, p. 343)

Following from this point, he states that any would-be theory of truth must imply an equivalence of the form:

(1) The sentence "snow is white" is true if, and only if, snow is white.

On the left side of the equivalence in (1), embedded in quotation marks, is the phrase "snow is white" and on the right side of the equivalence is the same phrase (snow is white), but here it is not in quotation marks. The significance of this difference, on Tarski's view, is that the right side represents the sentence itself, whereas the left side names the sentence. Generalising the procedure, Tarski arrives at the following:

(2) (T) X is true iff p.

Here, '(T)' refers to 'equivalence of the form', 'X' refers to a name of a sentence, 'iff' refers to 'if, and only if, and 'p' refers to (and is replaced by) the sentence named by X, to which the phrase 'is true' is predicated. This generalised procedure is the precondition of material adequacy that must hold for a theory of truth to be successful. Building on this formulation,

Tarski makes clear that he is interested in explicating the notion of truth in ordinary sentences, as a property of a sentence, but he also observes that sentences are only true relative to a particular language (e.g. (1) is true in English, but is meaningless in Italian or German etc.). Therefore, the analysis of truth must be relative to the object language (L) in question, which is formularised as follows:

(3) (T) X is true in L if, and only if, p.

This constraint of material adequacy shaped the form of the intricate theory of truth that Tarski went on to develop, but for the purposes of the argument in this chapter, this level of complexity is not required, for the constraint formulated in (3), hereafter referred to as a 'T-sentence', is adequate for current purposes, that is, when considered in conjunction with another aspect of Tarski's theory, his point of redundancy; that predicating 'is true' on a sentence adds nothing to the semantic content of the sentence.

Since predicating 'is true' does not add substantive content to the sentence, it can be argued that truth is not something external to the sentence, but rather, is infused into the structure of the sentence. It is on this structural point where the relevance of this theory to the self and PLTb starts to become apparent. Part of the beauty of Tarski's theory is in its simplicity and deflationary nature, always attempting to keep his theorising to an absolute minimum. In this sense, Tarski's theory of truth is consistent with the utterance aspect of PLTb theory. Specifically, both perspectives exploit structural configurations, as opposed to any concept with a substantive lexical content. Both perspectives link abstract nouns (e.g. 'self' and 'truth') to concrete, triangulable behaviours, rather than isolating the terms and performing conceptual analysis. In Tarksi's case, truth is part of a T-sentence structure and once a T-sentence is uttered, it can be classified as a behaviour and thus is susceptible to PLTb. The importance of this point, in connection to utterances, self-referring and PLTb, will now be further expanded upon.

Any linguistic utterance that is articulated with a communicative intent explicitly or implicitly acknowledges a distinction between the speaker and the receiver. Take Tarski's sentence 'snow is white', if this sentence is uttered by an individual with the intent of communicating a propositional meaning, then minimally there must be a speaker and receiver of the proposition for the communicative act to be (potentially) successful. This is true even if

the speaker is talking to him/herself, in which case the person would be both speaker and receiver. Take the context in which a person (P1) is asking another person (P2) about the colour of snow. The conversation may unfold as follows:

P1: What colour is snow?

P2: Snow is white.

This is a perfectly acceptable exchange, but within these sentences there is no word or words that indicate grammatical person, namely, who is the speaker (1st person), who is being spoken to (the intended receiver(s), 2nd person) and who or what is being spoken about (3rd person); things other than the speaker and intended receiver(s). In this instance, and in most instances of language use, grammatical person and intentions are signified implicitly as part of the social context of language use. The speaker expresses their thought in the form of an utterance, which is presumably thought in the 1st person. This utterance is addressed to someone that is the 2nd person, from the speaker's perspective, and finally, the utterance is about something identified in the 3rd person. It is possible that this implicit attribution or 1st and 2nd person could also be represented at the surface level of an utterance, without much altering the sentential and propositional meaning:

P1: I say to you: what colour is snow?

P2: I say to you: snow is white.

In fact, it could be argued that the clause 'I say to you' could be attached to any utterance in a language (here, English). For example, 'Hello' becomes 'I say to you: Hello', 'I love pasta' becomes 'I say to you: I love pasta' and so on. Generalising this theoretical procedure, we could arrive at the following formulation:

(4) 'U' ≡ I say to you: U

Where 'U' on the left side of the equivalence refers to a named utterance (i.e. the utterance mentioned), ≡ means 'is equal to' and the U on the right side of the equivalence refers to the

utterance in use, that which is mentioned. Hence, like in Tarski's T-sentence, there is a use/mention distinction. Exceptions to this procedure are those utterances where this grammatical person distinction is already part of the utterance at the surface level e.g. 'I say to you: No thanks!' would become 'I say to you: I say to you: No thanks!' This scenario occurs infrequently, for sentences of this type would be difficult to interpret.

As stated above, the reason why the clause 'I say to you' is rarely explicitly included as part of an utterance is because it is part of the agreed contextual understanding surrounding an utterance (i.e. is signified as part of the context). Thus, much like Tarski's predicate: 'is true', the clause: 'I say to you' is largely redundant in terms of substantive meaning contribution. In sum, consistent with Tarski's formulation 'is true', it can be said that the introduction of the clause 'I say to you' adds nothing to the semantic content of the utterance/sentence, for often this information is implicit in the utterance, saturated by the contextual environment.

In its current form, the above is a theoretical argument only, but there is empirical evidence that can be advanced to support this position. For example, on the basis of empirical data from linguistics, Ross (1970) argued that declarative sentences must be analysed as implicit performatives.

Building on the work of Austin (1962), Ross took performative sentences to require, by necessity, 1st person subjects (a speaker), 2nd person object(s) (an addressee) and a performative verb. More explicitly, he argued that all declarative sentences possess an implicit performative clause in their deep structure, where 'deep structure', in Chomskyan linguistics, refers to context free, syntactic derivation trees, which are transformed on the basis of certain rules (tree rewriting operations) into surface structure (i.e. that present in ordinary language). For example, the declarative sentence 'prices slumped', on Ross' view, becomes '*I pv you* prices slumped'; where *pv* represents a performative verb such as: 'say', 'think', 'demand' etc… and '*I*' and '*you*' are the necessary 1st person subject and 2nd person object. Holmberg (2010) concisely captures the essence of Ross' performative hypothesis and the types of linguistic data Ross summons for the existence of a higher 1st person subject 'I', a performative verb and a 2nd person object 'you'. Specifically, Holmberg focuses on Ross' arguments on the possibility of an anaphor to refer to a speaker in a main clause, without an overt 1st person antecedent. For example, consider the following sentences, paraphrased from Holmberg:

(5) Philosophers like myself/*himself often make mistakes

(6) John said that philosophers like himself often make mistakes

(7) An acquaintance is ring-fenced for a promotion.

In (5), the reflexive pronoun 'himself' is not permitted to function anaphorically for there is no appropriate antecedent, but it is permitted in (6) where 'himself' is embedded under a clause (i.e. 'John said'); it is 'John' in this clause that serves as the antecedent to the anaphor, thus making the anaphoric reference 'himself' grammatically acceptable in (6). Building on this point, sentence (7) is only interpretable if 'An acquaintance' is taken to mean 'An acquaintance of mine'. Therefore, on the basis of (5) and (6) – the requirement of an antecedent for correct anaphoric function – we can infer that (7) must be embedded under a covert clause (in deep structure) with a 1st person as subject ('I'), so that the pronoun 'mine' can function as a grammatically viable reference. This is one example of the type of linguistic evidence Ross develops for his performative hypothesis (ten empirical/linguistic arguments were outlined in total), which states, amongst other things, that there is an 'I' represented in a sentence, either overtly or covertly.

Ross' performative hypothesis covers a broad range of sentences and linguistic evidence, but at bottom, postulates a significant amount of invisible deep structure, thus, has faced stiff criticism. Nevertheless, though this level of linguistic representation is vulnerable to critique and thus has fallen out of favour, it does not detract away from the validity of the empirical observations made by Ross, for it is common in linguistics for formal frameworks of description to change. Yet, the general insight that speech acts have a syntactic dimension is a very active area of research and is arguably developing as the mainstream view (see Speas & Tenny, 2003; Wiltschko, 2017a; Wiltschko, 2017b). That said, even if one is not convinced that self-reference is part of sentence structure – either overtly or covertly, based on this empirical evidence from linguistics – one can find alternative means of supporting this point through a more general examination of languages other than English, hereafter discussed.

The focus so far has been on the grammatical first person in English, lexicalised in the word 'I'. However, in many languages, referring to the grammatical first person is achieved without a lexicalised pronoun. In generative grammar – that is, a linguistic theory that regards grammar as a precisely formulated set of rules that generates all of the grammatical sentences in a language – the technical term for a language that allows for this type of pronoun omission

is called a 'pro-drop language'. Most romance languages (languages that developed from Latin) are categorised as pro-drop languages with respect to subject pronouns. In these pro-drop languages, the expression of the subject pronoun is unnecessary (besides for emphasis), because person and number can be marked on the verb, facilitated by a highly inflected verb morphology present in the grammar. For example, in Vulgar Latin (the non-classical, common use of Latin) person distinctions can be marked on the verb (highlighted in red, below) which, in this instance, modify the verb stem 'am':

(8) Amo = I love

(9) Amas = you love

(10) Amat = he loves

Modern Romance verbs work this way also:

	I sing	**We sing**	**You sing**	**You all sing**	**He sings**	**They sing**
Catalan	canto	cantem	cantes	cantau	canta	canten
Italian	canto	cantiamo	canti	cantate	canta	cantano
Spanish	canto	cantamos	cantas	cantáis	canta	cantan

Table 3: Example of pro-drop languages with person distinctions (red) marked on the verb.

The important point to derive from this cursory look at the cross-linguistic evidence is that self-referring is not exclusively a lexical phenomenon, relating to the existence of a certain pronoun. Rather, self-reference can manifest in other ways, such as marking on the verb. This fact supports the general movement away from individual words (such as 'I') considered in isolation, towards a focus on structural, grammatical elements of an utterance.

To summarise the points made in this section and in §4.3, it was argued that the use of 'I', as a self-reference, might serve as an adequate behaviour to link to the profile of 'self', as part of a PLTb solution to theorising on this abstract noun. After this suggestion, several arguments were presented that indicated that the isolated use of 'I' might not sufficiently capture the phenomena of self-referring. It was pointed out that 'I' can only be understood as a meaningful self-reference when it coheres with the above-outlined requirements of normativity and statistical typicality, which can only be assessed at the utterance level (i.e.

where ‘I’ is part of utterance). If an utterance coheres with the requirements of normativity and statistical typicality, jointly referred to as ‘mid95%’, then the utterance can be said to be a functional move in a language game. Stated another way, language is an intricately-played public game, where the minimal constituent piece of this game is a normatively correct and statistical typical utterance, which itself consists of words such as the self-referring pronoun ‘I’. In this sense –slightly adjusting a previously used metaphor – an utterance can be conceived as similar to a chess move. For a chess move to be meaningful, it must occur in the context of a game of chess and likewise, for an utterance to be meaningful, it must occur in the context of the language game, which is equivalent to saying that an utterance must satisfy the mid95% requirement. From this conclusion, the discussion moved onto the more specific role of self-reference in mid95% utterances. To wit, it was argued that ‘I’ (or self-reference, more generally) might be part of the structure of an utterance, even when it is not overtly stated via a lexical pronoun. This idea involved an examination of Tarski’s theory of truth, which inspired the development of similar theory with respect to self-reference. This application of Tarskian theory to self-reference, as an aspect of the PLTb solution to the self, was then supplemented with empirical evidence from linguistics, namely, Ross’ performative hypothesis and evidence from pro-drop languages.

In short, it can be concluded that the lexicalised pronoun ‘I’ is not (or is not always) a satisfactory behavioural indicator of selfhood and thus, cannot be the locus behaviour of the PLTb solution to the self. Instead, the PLTb solution will be found at the utterance level, for it is here where the requirements of normativity and statistical typicality can be assessed. Not all utterances possess a self-reference at the surface level, but as seen above there are arguments that can be made to suggest that even if a self-reference (specifically, ‘I’ in English) is not present at the surface level, it may well be present at a different or deeper level. However, it might be the case that one is not convinced by arguments for self-reference (an ‘I’) present in utterance structure, nevertheless this does not detract value from the idea that a normatively and statistically typical utterance may serve as an optimal behaviour for the PLTb theory of self. In other words, even if ‘I’ is not present in an utterance (at surface, deep or other level) an utterance may still serve as an adequate behaviour to anchor to the PLTb profile of self. In the following section, utterances are further examined, to establish which type of utterances are indicative of typical selfhood, and thus, by default, the dimension of non-typical utterances and diminished selfhoods are also considered.

4.5 Utterances as self-behaviour

The ability to generate and understand utterances (i.e. the capacity for human language) is typically held to be a uniquely human characteristic, the quintessential human trait, a product of a genetic endowment of the central nervous system (CNS), peppered with linguistic experience in the world. However, it should be noted that whilst this proposition is accepted across the board, there is significant debate on the precise nature of the genetic endowment (i.e. is it a broad, domain-general mechanism or a domain-specific language mechanism?) and the required level of linguistic experience in the world in order for the capacity to fully form (see Dąbrowska, 2015 for a recent overview of the arguments). Fortunately, the arguments presented in this thesis do not hinge on the specific outcomes of these debates, for the arguments developed here, bootstrap from the assumption outlined above: that there is a genetic component to the acquisition and full development of language, that emerges from, and correlates with, processes in the CNS (predominantly, in the brain). Furthermore, this genetic component requires linguistic experience in the environment to develop to its fullest extent. Holding to this relatively high-level assumption avoids much of the controversy associated with the specifics of the mechanism(s).

In terms of the form that human utterances take, they too can be modelled using the normal distribution curve, as represented in Figure 4, below.

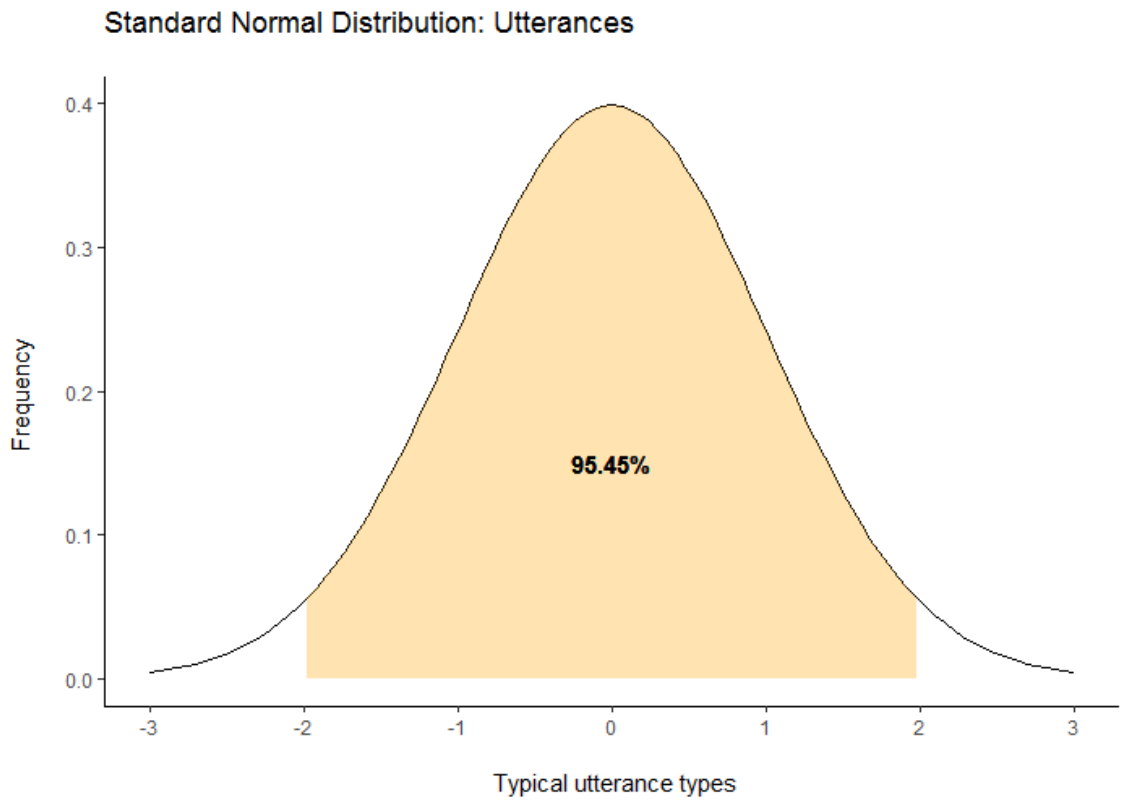

Figure 4: Simulated standard normal distribution curve to demonstrate the typical form of human utterances.

Similar to that described above with respect to 'I', utterances falling within the mid95% are of a certain kind. At a high level, they can be characterised as fully socialised and statistically typical, but in a more specific sense, mid95% utterances (i.e. those that dominate ordinary language exchanges) can be described as syntactically and semantically congruent/comprehensible/correct/acceptable (terms used interchangeably, to ease the flow of writing). In contrast to these, end5% utterances violate these mid95% principles, thus are semantically and syntactically incongruent/incomprehensible/incorrect/unacceptable. These two components, syntax and semantics, refer to the set of rules governing the structure of utterances and the meaning of utterances, respectively. Considered together, these two categories cover the two main facets of an utterance's constitution. Especially if semantics is taken to refer to meaning in the broad sense of the term, that is, to include the terminal yield of all aspects of grammar relating to meaning, such as pragmatics; there are strong neuroscientific arguments to suggest that semantics and pragmatics are intimately linked (see Dudschig, Maienborn, & Kaup, 2016; Hagoort, Hald, Bastiaansen, & Petersson, 2004). Likewise, for syntax, which is here intended to be understood in the broad sense, to include all aspects of

grammar relating to structural, combinatorial rules (e.g. including morphology and phonology). This dichotomy is quite a coarse characterisation of language but is adequate for current purposes.

An utterance is syntactically correct if it conforms to the combinatorial rules of the language in question. Syntactically correct utterances make up the significant majority of utterances produced in a language (mid95%) and are processed with absolute cognitive ease. In contrast to these, syntactically incongruent utterances (end5%) are easy to spot for they are cognitively jarring for a listener/reader to understand and for a speaker to generate in speech. This jarring effect can be caused by a violation of syntax or by an increased level of syntactic complexity present in the utterance. For example, consider the following sentences:

(11) Furiously sleep ideas green colourless (Chomsky, 1957/2002).
(12) The rat the cat the dog bit chased escaped (Kempen & Vosse, 1989).

Utterance (11) involves a string of syntactic violations. Whereas, in (12), the syntax is technically acceptable, but is particularly complex. In both cases, the utterances should prove somewhat jarring, to produce and comprehend; impossibly so for the comprehension of (11). The conjunction of this phenomenon with its low prevalence in ordinary language, is suggestive of an end5% type of utterance. In addition to the above method of identification (i.e. via the jarring effect), these types of utterances can also be identified through linguistic analysis, as briefly discussed in §4.3.

An utterance is semantically congruent (mid95%) if its content is understandable to other speakers and its content is understood as intended. Whereas, semantically incongruent (end5%) utterances are those which are not understandable to other speakers, these types of utterances are typically less easy to spot than the above outlined syntactic violations. Take Chomsky's most famous sentence:

(13) 'Colourless green ideas sleep furiously' (Chomsky, 1957/2002).

Here, (13) represents a sentence that is syntactically correct, but which does not have an easily decipherable meaning (i.e. its meaning is largely incomprehensible). Other, more veiled examples include utterances were the meaning is somewhat comprehensible, but the content is

reliably inconsistent with known aspects reality, such as in delusional utterances produced in some forms of schizophrenia. For example, consider the following exchange below between a therapist and a patient (Anna) with a diagnosis of schizophrenia, taken from Zimmerer, Watson, Turkington, Ferrier, and Hinzen (2017, pp. 3-4). Here, the therapist is inquiring about what the patient calls 'sight and mind painting':

Therapist:	What does the sight and mind painting involve?
Anna:	Ah, you can do it on the camera, on the chair and the top of a … a door, say, yes, flown in by radar.
Therapist:	So you put together a composition
Anna:	Yes
Therapist:	And its passed by radar.
Anna:	Yes
Therapist:	And… how does the radar work? Is there a kind of equipment?
Anna:	Yeah, yeah, I'm not sure about that.
Therapist:	So how, how do you process it, you've got the composition there…
Anna:	There are, there have canvasses, down in London and somehow, I see what I am looking at, that becomes a painting. The painting is then transferred.
Therapist:	The canvasses down in London, somehow your composition is transferred there?

On a kind, but still very uncertain reading of the above conversation, Anna's 'sight and mind painting' could be interpreted as some form of art that is performed on several objects (e.g. a camera, chair, top of a door), which gets transferred to London, by radar, and which manifests as a painting on a canvas. Here, Anna's utterances are more-or-less, syntactically acceptable, as she is approximately following the combinatorial rules of English, but the meaning and the import of the utterances (i.e. their propositional content) are highly unusual and largely disconnected from reality and its physical/technological limitations. In certain circumstances, these types of utterances are referred to in psychiatry as delusional or as delusional beliefs (American Psychiatric Association, 2013, p. 87). Delusional beliefs, which can only find their expression in an utterance, are beliefs that are clearly false (Kiran & Chaudhury, 2009), to the

extent that they indicate an underlying cognitive, neurological dysfunction; an end5% neural architecture and resultant cognitive phenotype (more on this below and in Chapter 5). Distinguishing a delusional belief from a workaday, overvalued, mistaken belief – which we all likely possess – can be difficult. However, the former is held with strong conviction, even in the face significant contrary evidence, whereas the latter usually allows for some reasonable level of doubt.

One might have noticed an asymmetry between syntactic violations and semantic violations, namely, that all syntactic violations are also semantic violations, but not all semantic violations necessitate that an utterance is also violated syntactically (e.g. as in (13); though with significant background context semantic meaning could be established). This means that end5% utterances can be further subcategorised as those which have incomprehensible meaning, but are syntactically correct, and those that possess violations of syntax, which by default also prohibits comprehensible meaning.

The argument here, then, is that utterances of the mid95% form may serve as a behavioural indicator of self, necessary to stabilise this term via PLTb. In this way, the term could potentially reach a high level of synchronicity across the linguistic community. If one accepts this proposition, that a mid95% specification for an utterance is an adequate behavioural indicator of self, then the production and comprehension of a mid95% utterance indicates an 'intact' selfhood.

4.6 Objections to PLTb theory

It could be argued against this PLTb position on several fronts. For example, one suggestion might be that mid95% utterances (and their end5% counterparts) do not capture the complexity of self associated with humans, and so, are a shallow means of assessing selfhood. This criticism is tending towards the metaphysical approach criticised in Chapters 1, 2 and 3, in that it appears to assume to know what the self *is*, which as discussed previously, is problematic. But, in a more direct response to this criticism, one could draw attention to the fact that the production (and comprehension, discussed below) of a syntactically and semantically congruent (mid95%) utterance requires the deployment of a large array of cognitive capacities, such as long and short-term memory, attention, theory of mind and cross-modal sensory perception, virtually everything humans possess goes into mid95% utterances. Therefore, utterances, far from being shallow, represent the workings and interactions of a highly-complex cognitive system. On this view, the proposal of mid95% utterances as a behavioural indicator

of self seems completely reasonable. However, on this response, a critic could reply by acknowledging that the production of an utterance requires a huge array of cognitive capacities, yet still maintain that utterance behaviour is a shallow means of capturing the self. For example, they might argue that an aphasic patient, one who has a damaged/limited capacity for speech production, may still be cognitively sharp in all other respects, and would unjustifiably (on their view) be characterised as possessing a diminished selfhood on the PLTb theory, for their utterance production is of the non-typical type (end5%). Consistent with this, they might also add that a pathological mass murderer, whose utterance processes are within the typical range (mid95%), would also (on their view) unjustifiably be classified as possessing the reverse, an intact selfhood. This criticism fails on two levels. Firstly, in suggesting that the above-outlined aphasic patient has an intact selfhood and that the murderer does not, the critic makes the mistake of bringing their own definition (of self) to bear on the debate. Presumably, their thought pattern unfolded as follows: they considered the implication of the PLTb profile in comparison to their own definition, and decided the former to be inadequate, hence the rebuttal. The problem here is that this approach requires that there are necessary and sufficient conditions for the self, which can be established via classical conceptual analysis, for how else could one call upon this criticism?

In addition to this, and to be clear, it is not the suggestion of this chapter that the PLTb theory of self, utilising utterance behaviour, constitutes a complete profile of selfhood. Rather, utterances are here isolated as an exemplar behaviour, a triangulable behaviour typically associated with selves, and thus worthy of consideration in the PLTb profile of self, under development. It is not the suggestion that the PLTb theory is complete; in fact, the contrary is likely the case, for a broad consensus is required on the important behaviours to anchor to the profile of self. It is not the aim of this thesis to develop the profile in its entirety. Rather, this thesis, and in particular the argument of this chapter, aims to motivate the uptake of an alternative approach to abstract concepts/nouns (i.e. via PLTb), which can be characterised as an alternative, metaphysical program to that advocated in the classical, analytical philosophical approach, which problematically radiates far beyond the discipline of philosophy.

One of the main criticisms of the classical metaphysical approach, developed in this thesis, is that it generates a set of positions that when considered individually seem plausible, but when considered as a complete set of theories give rise to wide scale contradiction and incompatibilities (see §1.6). On this point, one might question whether this PLTb approach also falls foul of this issue. This possibility is unlikely for several reasons. Firstly, the

contradiction associated with the classical approach springs from the postulation of essential qualities (of the self), which in turn leads to reification and quasi-tautologies (as detailed in §3.5 and §3.6). This is not a problem for PLTb, for it is not wed to the existence of essential cores and thus, is not trapped by these reification biases. Instead, PLTb leans on established methods in science, particularly those found in the scientific approach to abstract, non-perceivable, theoretical entities (as described in §4.2). Therefore, the PLTb approach is bootstrapping from a completely distinct set of assumptions. Second to this, an important aspect of this PLTb approach, not yet overtly stated, is the value it confers on the seeking of consensual agreement amongst theorists in forming the best behavioural profile for the self in humans. A collaborative effort to make headway with abstract concepts, as opposed to the purely individualistic enterprise to find the absolute, essential truth on the matter, which can often encourage a competitive and uncooperative spirit within a discipline. If a relatively strong profile were to emerge on the self-concept, via PLTb, then it could be openly factored into important philosophical debates that utilise this concept (e.g. euthanasia, abortion, advanced directives and so on). However, even if the approach were extensively adopted, it is still unlikely that there will be complete agreement on the behavioural profile, for individuals will come to the problem from different backgrounds, with different belief systems, and thus, may value and advocate different behaviours to different extents. In a similar manner, physicists, working within different scientific paradigms (Kuhn, 1970), may have differing perspectives on the fundamental behaviours attributable to abstract theoretical entities. Nevertheless, if this scientific analogy and above argumentation holds true, the level of disagreement in PLTb approaches will be significantly less than that associated with the classical, analytical, metaphysical approach to concepts, because the concept will be accountable to realities of the empirical world, manifest in behaviours.

Another criticism might be that there is a slippery slope, thin edge of the wedge, concern to PLTb, namely, that the adoption of this PLTb approach might lead to the prohibition of certain content in an utterance. Stated in other words, might the application of PLTb stifle creative thought or controversial ideas by ostracising those who operate outside this norm (i.e. those individuals who utter end%5 content)? This latter idea is based on the point that the production of utterances of a certain form (mid95%) are here interpreted as an indication of an intact selfhood, whereas production of utterances outside of this form (end5%) are seen as an indication of an attenuated or diminished form of selfhood. Though this PLTb approach speaks of norms and makes the above-mentioned demarcation, it does so from a statistical perspective

and, as such, is free from moral judgment and value attribution. Put differently, this PLTb approach is, at-heart, a value-free statistical model. It is not the case the mid95% utterances are good whilst end5% utterances are bad, rather it is simply that the former are statistically typical whereas the latter are statistically non-typical; remembering that the motivation for this type of approach is to help stabilise abstract nouns by anchoring them to common and relevant behaviours. On the first point, of stifling creativity and/or controversial ideas, applying the PLTb method to utterances does not lead to prohibition, for reasons outlined above, but also, technically speaking, these types of utterances (i.e. creative/controversial) would not likely be of the end5% specification. To reiterate, end5% utterances are those that are syntactically and/or semantically incongruent. On the end5% specification, creative/controversial ideas, though perhaps not impossible, are at least very unlikely to arise. For an utterance to be creative/controversial it must be meaningful and thus, must also be syntactically correct, so by default, will be part of the mid95%. But more than this, for an utterance to be controversial, requires that it be contrary to some currently held norm or norms, which, in itself, demonstrates a high-level knowledge of the language game in play. In short, the mid95%-end5% distinction does not infringe on speech freedom and neither does it attribute moral value on those individuals producing either of these types of utterances or comprehending utterances in an end5% fashion.

Continuing in this dialectical approach to possible objections, PLTb theory might also be accused of not being exclusive enough to humans. For example, if a computer/machine, operating on machine-learning principles (e.g. neural networks), were able produce a set of utterances that would pass the Turing test, would this mean that they have the same form of self associated with humans? The short answer to this is 'possibly', but it depends on whether the computer in question has the same level of understanding, associated with their utterances, as humans do.

The Turing test, developed by Turing (1950), is held as the gold standard for assessing the ability of a machine to simulate human intelligence. The test involves a machine producing linguistic output (behaviour) during a conversation with a human, to see whether an additional person (the judge) can distinguish between the two (i.e. accurately identify which respondent is the machine and which is the human). The judge is aware that one of the two in the conversation is a machine designed to produce human-like responses but is only able to make their judgment based on the conversational/linguistic output. The conversation is presented to the judge as text on a computer screen, so that the computer does not have to render their

response in audible speech, which, at the current advance of technological, would immediately give the game away. For a machine to pass the Turing test, requires that the judge not be able to accurately determine which of the two participants is human and which is the machine.

It can be argued – but will be argued against, below – that once a machine passes the Turing test, it will satisfy the condition of artificial intelligence (AI), a level of intelligence approximately equivalent to humans. Furthermore, because the Turing test measures intelligence through language, it can also be argued that if a machine passes, it can be said to be producing utterances of a comparable form to humans, hence the earlier question: would an AI machine, on the PLTb approach, have the same form of human self? However, this question can only be answered conclusively when it is clear what level of understanding/comprehension the AI is experiencing during the utterance production and comprehension. This information is important for reasons outlined in Searle's (1980) 'Chinese Room' thought experiment, in which, Searle imagines a person (P) in a room full of Chinese characters and a book of instructions on how to respond to Chinese messages; P cannot understand the Chinese characters but can understand the book of instructions. Now, if a Chinese speaker (CS), outside the room, pushes Chinese messages under the door to P, P can follow the book of instructions in selecting an appropriate response to push back under the door to CS. In this scenario, CS would believe that s/he is communicating with a Chinese speaker, while in reality they are communicating with a person (P) that does not understand Chinese, but who does know how to follow a book of instructions on using Chinese characters. This thought experiment serves to highlight the point, by analogy, that even if an AI can convince a judge that it is human (i.e. pass the Turing test), it does not necessarily follow that the AI has mastered language to the same extent as humans, for it may be that it is merely following a set of instructions (computer algorithms) and thus is devoid of the accompanying understanding that humans possess. On this conception of AI, the linguistic behaviour would be a mere simulation of human utterances and would not qualify for the same form of selfhood. However, if an AI mastered language to the same extent as a human (including understanding) its utterances would qualify as human equivalent (i.e. mid95% specification), so would justify the attribution of the same form of self. In fact, this point extends to anything that masters language to this extent, this includes the possibility of discovering this ability in other non-human animals or extra-terrestrial life forms.

This requirement of understanding leads to an important follow-up question, namely, 'what is the nature of the human understanding/comprehension process, for this has not been discussed so far?' Or alternatively, 'what is the mid95% profile for human utterance

comprehension?' The answer to this question lies in a behaviour other than that associated with the production of utterances (discussed above), for these are manifest in speech output, with the speech sound functioning as the behaviour, triangulated and assessed for normativity and statistical typicality. This method of approach is unavailable for comprehension, since comprehension – at least, at one level of understanding – is an internal process, primarily of the brain; thus has no obvious external behavioural presence. Therefore, one must find another way to triangulate this aspect of an utterance. But given the broad use of the term 'behaviour' employed in this thesis, one can find a means of triangulating this process relatively easily, for we can apply PLTb to neuronal activity/behaviour associated with comprehension. In order words, we can examine/triangulate neuronal behaviour during the comprehension process of language and develop a mid95% and end5% neuronal profile based on these scientific observations.

The relation between the cognitive function of comprehension and the underlying brain activity that correlates and facilitates this function, has been studied through various techniques in cognitive neuroscience, predominantly: electroencephalography (EEG), magnetoencephalography (MEG) and functional magnetic resonance imaging (fMRI). All of these methods are viable approaches, but for reasons that will become apparent in the next chapter, the focus of this thesis is on EEG, specifically, the event related potential (ERP) components that relate to syntactic and semantic processing, the N400 and P600, respectively. The specific details of these components are theoretically and experimentally explored in Chapter 5, but for now it can simply be stated that the N400 and the P600 are triangulable behaviours that can be indexed in terms of mid95% and end5% specifications. Thus, once fully explored, will serve as the comprehension aspect/behaviour of the utterance profile in development here, with regards the self.

Considering this approach to comprehension, one might wonder why it has not been suggested to integrate the neuronal behaviour underpinning speech production into the profile of utterances, as a further proxy-behaviour of self. Theoretically speaking, it is a sound suggestion to integrate this and all other relevant information into the self-profile, to make this profile more precise, but for two practical reasons, this is not possible in this thesis. Firstly, whilst there is a growing body of neuroimaging evidence on language production, the currently existing neuroimaging techniques are severely limited in their examination of language production, due to motion-induced artefacts associated with producing speech, which interfere with neuroimaging recordings (Birn, Bandettini, Cox, & Shaker, 1999; Fiez, 2001; Gracco,

Tremblay, & Pike, 2005). Secondly, as touched upon above, this thesis has a finite scope so it is not possible to pursue the self profile from every possible angle. Instead, the theoretical work that has been developed so far, and the empirical elements that will follow, constitute a first attempt to motivate a different approach to theorising, which emphasises the building of PLTb profiles for abstract nouns/concepts. The specific behaviour considered in this chapter is the utterance, as part of the wider human-language faculty, but again, as stated above, this is not to say that other behaviours are not important in the profile of self, for they most certainly are. Rather, utterances are here advanced as an exemplar behaviour, to motivate and kick-start this new programmatic approach.

4.7 Chapter conclusion

It is all too easy to criticise a theoretical position, to find holes in its arguments and to compute unusual and unwanted consequences from its premises and conclusions. The real intellectual challenge comes when one tries to replace the thing criticised with something of increased durability, with a theoretical perspective that better captures the object of enquiry. This chapter can be regarded as an attempt to do this, for the preceding arguments – those developed in Chapters 1, 2 and 3 – were aimed (in part) at critiquing the classical, metaphysical, essentialist approach to concepts; specifically, those applied (consciously or unconsciously) to the self. Whereas this chapter, following on from this critique, is an attempt to replace that which has been dismantled, with a new, positive way of theorising on abstract nouns/concepts, namely, by linking them to triangulable behaviours. The appeal to behaviour is advanced as an antidote to classical, essentialist metaphysics.

This chapter began by outlining why abstract nouns require triangulation to be useful across a linguistic community. It was shown that PLT is not a workable candidate for these terms, for they are devoid of a clear third-point physical object, but that PLTb may be a suitable alternative, since it has the necessary triangulation feature and has proven useful in the scientific examination of abstract phenomena. In consideration of this point, PLTb was then applied to the concept of self. Specifically, this involved the building of a behavioural profile indicative of selfhood. In the first instance, the behaviour of self-referring, via the lexical pronoun ‘I’, was considered as a possible behavioural indicator of self. This was chosen, for it is a uniquely human, linguistic behaviour, a defining characteristic of human communication, which supports and facilitates a range of cognitive capabilities. Furthermore, self-referring via

the pronoun ‘I’ has received a great deal of attention in the literature associated with the self and is often held as a crucial aspect of the concept.

However, after careful consideration of the pronoun ‘I’, it was established that it would not serve as an adequate behaviour in the profile of self, but that statistically typical, fully socialised, utterances would. After reaching this conclusion, the details of how these utterances satisfy the requirements of normativity and typicality were examined. It was argued that utterances can be categorised as behavioural indicators of an intact selfhood based on a linguistic examination of their syntactic and semantic congruency, in the context of the Gaussian normal distribution model.

On the completion of the above point, PLTb theory had been described and its application considered with respect the self, i.e. using utterances as a behavioural-proxy of selfhood. With the position in place, the argumentation then switched to possible objections. These included the accusation of shallowness, the classical incompatibility problem, the thin edge of the wedge criticism and the AI argument, the latter of which highlighted the importance of human comprehension of an utterance. In consideration of this point, in the final stages of the chapter, it was proposed that human comprehension be integrated into the PLTb profile of self via the triangulation of neuronal behaviour associated with the N400 and P600, EEG ERP components. Current evidence suggests that these components index semantic and syntactic comprehension, respectively. In the next chapter, this line of argumentation will be further developed, marking the beginning of a shift in thesis-focus from theoretical to experimental.

5 TWO EEG LANGUAGE EXPERIMENTS

5.1 Chapter Overview

In the previous chapter, a new method for approaching abstract concepts was proposed, in which the methodological emphasis was on the building of behavioural profiles, as a means of stabilising abstract concepts, so that they may be utilised and communicated effectively. This method was developed with reference to the abstract concept 'self', with a specific focus on language production behaviour. However, it was noted that a crucial aspect of language was still not accounted for in this profile, namely, language comprehension. To remedy this, two new electroencephalography (EEG) experiments are here presented which examine the neuronal processes underpinning the language comprehension mechanism.

The first experiment (§5.6) is with neurotypical adults and examines the N400 and P600 event related potentials (ERPs), which are associated with the processing of meaning and syntax, respectively. The second experiment (§5.7) utilises the same experimental paradigm, but with patients with a diagnosis of schizophrenia and neurotypical controls. This second experiment is a pilot experiment, the aim of which is to perform an initial exploration into possible physiological differences between patients and controls, and to assess the feasibility and acceptability of the design for use with patients, as a potential precursor to future, full-scale, patient EEG projects (see §5.7.1 for an outline of other pilot aims). Patients with a diagnosis of schizophrenia were chosen, because schizophrenia is often linked to a loss of self or attenuated selfhood (Kircher & David, 2003; Sass & Parnas, 2003). Therefore, it was of interest to explore to what extent this is the case, considering the newly developed PLTb approach to the self.

To give these experiments the necessary background context, this chapter begins with an overview of the EEG technique (§5.2), followed by a brief review of the EEG literature examining language comprehension, specifically that discussed with reference to the N400 and P600 ERPs. Finally, in §5.10, the empirical evidence derived from the two experiments – considered in conjunction with the previous EEG literature on language comprehension – is discussed as a possible form of behaviour to integrate in the PLTb profile of self, and what this would mean for patients with a diagnosis of schizophrenia.

5.2 The EEG technique

The processes underpinning language comprehension are complex and unfold rapidly. Therefore, in order to examine these processes – with a mind to integrating this information into a PLTb profile – a method of research is required that has a suitably high temporal

resolution; that is, if the research question is aimed at probing the fine-grained temporal aspects of these processes. One such method which satisfies this criterion and which has played an instrumental role in the understanding of language processes is the recording of electrical brain activity through the EEG imaging technique.

The EEG signal, first discovered by Hans Berger in the 1920's (Millett, 2001), represents a non-invasive, yet, direct and continuous measure of cortical brain activity. Specifically, EEG measures fluctuations in voltage resulting from ionic current processes within the neurons of the brain, the dominant explanation being that this signal arises from the post-synaptic potentials (PSPs) produced by populations of cortical pyramidal neurons that fire in synchrony (Kirschstein & Köhling, 2009). These cells, with their long apical dendrites, are the likely main contributors to the EEG signal, because of their proximal and perpendicular layering, close to the cortical surface, and because of their commonly orientated direction, which allows synchronic summation of neuronal activity. Action potentials are a less likely source, because they are shorter in duration than PSPs, and are therefore less likely to fire synchronously (Nunez & Srinivasan, 2006).

Many EEG language-processing studies examine event related potentials (ERPs). ERPs are time-locked voltages that are embedded within the continuous EEG signal. These ERPs are time-locked to an event of interest and extracted for analysis; on the assumption that a particular ERP component represents the brain's response to a particular cognitive, motor or perceptual event (Luck, 2005). Typically, ERP components are referred to, first by a letter, referring to their polarity, either P or N, for positive or negative, followed by a number which either indicates the latency of the ERP, in milliseconds relative to stimulus/event onset, or by the ERPs ordinal position in the electromagnetic waveform. An ERP is made up of components of activity arising from electrical sources within the brain that are temporally and spatially distinct. Previous research has identified these components and their neural/cognitive underpinnings.

In addition to ERP analysis, there has been a recent influx of language processing studies that utilise a technique called: frequency analysis (Bastiaansen & Hagoort, 2015; Lam et al., 2016), in which the continuous EEG signal is decomposed into multiple frequency bands: Delta (1-4Hz), Theta (4-8Hz), Alpha (8-12Hz), Beta (15-30Hz), Gamma (30-90Hz) and (sometimes) High Gamma (>50Hz). Here, the different frequency bands are said to reflect different neural networks underpinning functional processes, again: cognitive, motor and/or

perceptual. This technique is not the focus of this thesis, but nevertheless, is a promising avenue of research in language processing, see §5.9 for future plans in this area.

The raw electrical activity from the cortex is recorded through silver/silver chloride electrodes (or sometimes, tin electrodes) held near to, or directly on, the scalp, usually via a tightly-fitted, elasticated cap. A conductive gel is then applied, to establish a connection between the scalp and electrode. This is followed by a light abrasion of the skin, to diminish electrical skin potentials and to establish a low impedance value between scalp and electrode (Kappenman & Luck, 2010).

There is some flexibility with respect to the number of electrodes required for recording, most often this decision is dictated by the needs of the research question and the degree of spatial resolution required. For example, in the typical language processing study, which targets broadly distributed ERPs such as the N400 and P600, an experimenter requires a minimum of 20 or 32 channel/electrode system. Higher density configurations, up to 256 channels/electrodes, are available and offer improved scalp topography and modelling of the underlying electrical sources. Unfortunately, in addition to this, they also increase setup time (substantially), increase the possibility of electrical 'bridging' between electrodes – where the conductive gel spreads between adjacent electrodes, distorting the EEG signal (Alschuler, Tenke, Bruder, & Kayser, 2014; Tenke & Kayser, 2001) – and increase the likelihood of detecting artefacts in a channel, resulting in an increased potential of data loss.

Once the number of electrodes has been decided, the next important technical/research decision is the choice of reference electrode. The voltages expressed in the EEG are relative measurements, reflecting the difference in electrical potential between two points/electrodes. The typical EEG system uses a ground electrode and a reference electrode to reduce noise common across all electrodes. In language processing studies, it is typically the case that the reference electrode is placed on the left or right mastoid, the thick bone structure behind both ears, as it is necessary that the reference electrode be as electrically neutral as possible; though in practice nowhere on the body is electrically neutral. During the offline analysis stage, a re-referencing process may occur in order that spatial distortion is minimised (Woodman, 2010, p. 11). As one example, a researcher might re-reference to the average signal activation of combined left and right mastoids. Additionally, though not common-practice in language processing studies, one may assign a global reference using an average signal across all electrode sites. The choice of reference critically affects the amplitude and location of electrical activity across the scalp, it is therefore necessary to ensure that the details of the selected

reference are outlined in the methods section of any published EEG research, in order that a comparison of results and replication of methods is possible. More generally, to ease this comparison and replication process, it is advisable to select the conventional reference-type used in the literature.

In addition to the electrodes placed on the scalp, further electrodes are often placed directly on the face, to help delineate between cortical activity and movement artefacts. Horizontal eye saccades and vertical blinks, for example, cause large voltage changes in the EEG signal; square-shaped positive waves in the former case and diphasic-like potentials of approximately 100 microvolts (μV) in the latter case, in comparison to 1-10μV typical for ERP components (Woodman, 2010). Facial electrodes on the outer canthus of each eye facilitate the recording of horizontal electro-oculograms (HEOG), which allows for correction of horizontal eye saccades, whereas electrodes placed on the infraorbital ridge(s) below one or both eyes and an electrode placed on the nazion (above the eyes), facilitate the recoding of vertical electro-oculograms (VEOG) which allow for blink correction. Note that preventive measures are also taken to reduce these forms of artefacts prior to experimentation. For example, the participant is instructed to relax as much as possible, to prevent general muscle artefacts, and the design of the experimental paradigm is such that it tends to involve a fixation cross that the participant is instructed to focus on, to reduce saccades. On the related note of artefact prevention, many researchers introduce a quasi-random jitter timing to the inter-trial stimulus, to ensure that frequency oscillations of the participant, particularly alpha-wave activity, do not become phase locked to the stimulus presentation rate (Woodman, 2010).

In the typical EEG lab, there will be two computers, one that is involved in the presentation of stimuli to the participant (hereafter, referred to as 'presentation computer'), and another that samples and stores the acquired EEG data (hereafter, referred to a 'data computer'). The presentation computer sends triggers to the data computer at specific times during the experiment, usually when stimuli are presented or when behaviour responses are made by the participant. These triggers enable offline extraction of the time-locked signals (i.e. ERP components) from the otherwise continuous EEG. The EEG signal is often minuscule and therefore requires substantial amplification, this is performed by the EEG system itself and is presented (live) on the data computer and stored as such. Recorded EEG data, in its raw unprocessed form, will contain electrical noise, such as that from any electrical devices near the participant. This noise is often manually attenuated, by shielding electrical equipment and/or having the participant perform the experiment in a faraday chamber.

As part of the typical ERP analysis pre-processing, the raw EEG data file is first visually inspected for gross artefacts that can be removed, preferably whilst blind to any of the experimental conditions. These artefacts include periods of recording that are not part of the experiment (i.e. before and after the task and during participant breaks) as well as large artefacts from participant movement. After this, filtering is performed on the data. Filtering involves the removal of certain frequencies from the EEG signal. High-pass filters let high frequencies pass whilst attenuating those frequencies lower than the set threshold. Low pass frequencies let low frequencies pass whilst attenuating those frequencies that are above the set threshold, and there are band-pass filters that combine high-pass filters and low-pass filters. High-pass filters are used to reduce slow drifts and skin potentials, whereas low pass filters are used to reduce muscle activity. Filtering is an extremely complex topic and if used incorrectly can lead to loss of information and a correlative distortion of signal (Luck, 2005).

Filtering is typically followed by one or more specific, usually semi-automated, artefact-correction procedures, such as blink removal (Croft & Barry, 2000) and ECG removal (Devuyst, Dutoit, Stenuit, Kerkhofs, & Stanus, 2008). Like filtering, these forms of artefact-correction aim to attenuate the presence of non-relevant electrical activity. Next, the continuous EEG file is segmented into the relevant trials (time locked to the event of interest). A baseline correction is then applied to each trial by subtracting the average voltage in the baseline window from all the data points in the epoch. Short baselines minimise overlap with previous trials/events, whereas long baselines increase the reliability of the baseline estimation. As an example, one baseline setting might span from -200ms prior to stimulus onset, at 0ms. Finally, the individual trials, as particular instances of certain conditions, are averaged together point-by-point for each condition, electrode and participant, to form a grand average for each experimental condition at each electrode. This information is then converted into numerical/spreadsheet form, ready for formal statistical analysis. In language ERP studies, this generally involves repeated measures ANOVA's and *post-hoc* dependent t-tests.

Audibly and visually presented words and sentences are the typical stimuli in language studies, but occasionally, pictures and videos are also utilised. This range of possibilities allows a researcher to investigate various questions relating to language processes, though, there are some restrictions and best practices that one should consider when designing a language-based EEG experiment. Firstly, in order to prevent horizontal eye saccades with visually presented sentences, most studies employ a technique in which the words of a sentence are presented one-at-a-time in the centre of a screen, scaled in terms of font to restrict movement in the visual

field. Moreover, in both audibly and visually presented experiments, there is usually a fixation point for the participants to focus on, in the former case this is presented throughout the experiment, whilst the participant listens to the stimuli, whereas in the latter case the fixation point is only presented in-between the presentation of words, to help maintain central focus.

Another consideration with language studies is the relatively large number of trials required to isolate language orientated ERP components, such as the N400 and P600, between 30-60, per participant and condition, according to Luck (2005) and Woodman (2010, p. 7). This requirement relates to the signal to noise ratio (i.e. the size of the ERP components relative to the background EEG activity). The smaller the signal, the more trials are required for the averaging to reduce the background noise to a point where the signal is clear. In language studies, it is also important to control for word frequency and word salience when possible, and finally, as is the case with all psychological experiments, full counterbalancing of stimuli, across all conditions, should be done, if appropriate.

5.3 The N400 ERP

The N400 ERP component, and its modulation during language comprehension, was first discovered by Kutas and Hillyard (1980), in which, they contrasted the ERPs elicited by congruent sentence endings (e.g. 'It was his first day at **work**') with those elicited by incongruent sentence endings (e.g. 'He spread the warm bread with **socks**'). In doing so, they found that the latter, the incongruent sentence ending, correlated with an increased negativity around 400ms after the anomalous word's onset (e.g. 'socks'), compared to the congruent sentence ending (e.g. 'work'), see Figure 5.

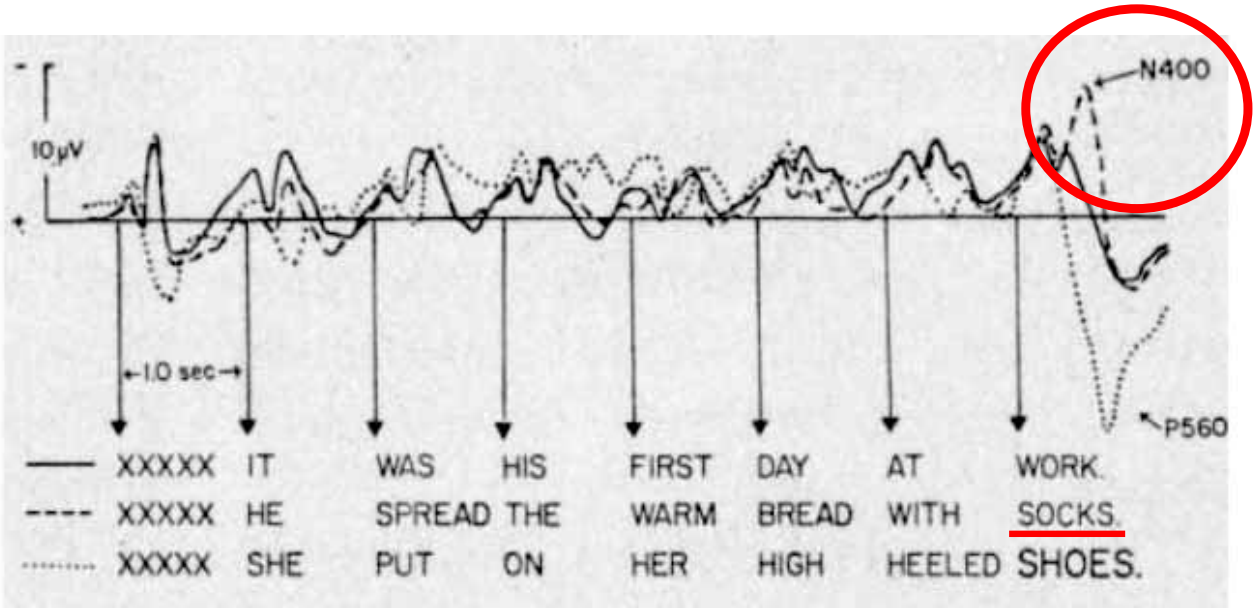

Figure 5: Grand EEG waveform taken from Kutas and Hillyard (1980). An increased N400 correlated with the semantically incongruent sentence ending 'socks' in comparison to the semantic congruent ending 'work'.

Following this discovery, this ERP component received (and continues to receive) significant attention in the neuroscientific examination of language comprehension; see Kutas and Federmeier (2011) for a 30-year review of research into this component.

Topographically speaking, the N400 ERP is a centro-parietally distributed negative waveform that tends to peak at approximately 400ms after stimulus onset. Comprehension experiments have demonstrated the N400 effect not only with spoken and written word stimuli presentation, but also with a broad range of modalities including sign language, drawings and pictures (Federmeier & Kutas, 2001; Kutas, Neville, & Holcomb, 1987; Ortu, 2012).

The amplitude of the N400 is modulated by a variety of factors, other than just semantic violations. One predictor of the N400 amplitude relates to semantic expectancy (Hagoort & Brown, 1994). More specifically, the N400 response to a word is inversely related to the word's cloze probability; where cloze probability is defined as to the degree to which the context establishes expectation, calculated as a proportion of participants who would provide the word in question as a continuation of a sentence fragment, when given offline in a behavioural task (Kutas & Federmeier, 2011). Stated in other terms, if the word meaning and the unfolding sentential context fit tightly together in terms of expectation and semantic-relatedness, the N400 response will be low, that is, relative to a word meaning in a context with a weaker, more unpredictable fit. Because of this observation, it is often stated that the N400

reflects the processing of integrating meaning into the linguistic context (Hagoort & van Berkum, 2007; Osterhout & Holcomb, 1992), with an increased N400 representing increased difficulty in integration.

However, comprehending an utterance is not simply a matter of integrating the lexical content associated with a word into a current unfolding linguistic content. For in all cases, these utterances are part of a language game, which is intimately related to the world. In fact, some of these utterances – the propositional ones – make truth claims about the world. Therefore, a theory of understanding and comprehension processes should connect with, and index, a speaker and listener's world knowledge. A physiological explanation that could not account for this facet of comprehension and understanding, would be as shallow as the theoretical account of comprehension and understanding exposed by Searle's Chinese room example (see §4.6), and would not be fit for purpose regards integration into the PLTb profile of self.

Broadly speaking, there are two basic approaches to explaining the deeper notion of understanding and comprehension discussed above. These are typically termed the 'one step (or one stage) theory' and the 'two step (or two stage) theory'. If one holds to the Fregean (1892), and later Fodorian (1998), view of language, as governed by principles of compositionality – that overall sentence meaning is a function of its word parts in conjunction with a set of syntactic rules (a position argued against in §2.4) – then by implication, it appears to follow that language comprehension unfolds in (minimally) a two stage fashion (Cutler & Clifton, 1999; Dudschig et al., 2016; Lattner & Friederici, 2003). In this first stage, a context-free computation of the individual and fixed word meanings is performed, the combination of which gives rise to the overall sentence meaning. Then, in the second stage, this literal sentential meaning is saturated with the prior knowledge of the speaker/listener (i.e. with their 'world knowledge'), part of which is the knowledge derived from the context in which the utterance was stated. As an example of this, take the sentence 'snow is blue'. On the above outlined, two-stage model of language comprehension, firstly, the individual word meanings of the sentence (e.g. 'snow', 'is' and 'blue') are sequentially computed to generate the overall meaning of the sentence, that is, in accordance with the syntactic rules that govern their combination. Then, once this sentential meaning is formulated, the other forms of information are integrated, including contextual information and speaker/listener world knowledge. On this view, it is in the latter stage where the listener's previous knowledge about the colour of snow is integrated, and it is only at this point where the sentence 'snow is blue' can be assessed for truth and falsity with respect to state of affairs in the world. Based on this theory, it would

appear to follow that different physiological processes are in play in the different stages, and that the N400 indexes the first, purely linguistic/lexical stage of information processing and that some other process accounts for integration of world knowledge.

A number of researchers have argued against this two-stage model of comprehension, stating that the distinction between lexical/word meaning and the world knowledge of the speaker/listeners is a false one, for word meaning is intimately linked to an individual's world knowledge, such that the former cannot be established in the absence of the latter (Hagoort et al., 2004; Hagoort & van Berkum, 2007; van Berkum, Zwitserlood, Hagoort, & Brown, 2003). As an alternative, these researchers argue for the assumption of immediacy, which holds that all aspects of meaning are processed immediately and simultaneously, and they attempt to demonstrate this point empirically.

One of the first EEG experiments exploring this immediacy assumption was performed by Hagoort et al. (2004), in which they investigated the differences between processing violations in linguistic information and violations in world knowledge (i.e. false utterances). In this experiment, participants were presented with three versions of the same sentence, modified only by the change of an adjective (in bold), such as: 'The Dutch trains are **yellow/white/sour** and very crowded', see Figure 6. The sentence with the adjective 'yellow' is considered true, for Dutch trains were known to be of this colour. Therefore, by default, the sentence with the adjective 'white' is false, in consideration of the truth of the first sentence, and thus is a world knowledge violation. Finally, the sentence with the adjective 'sour' represents a semantic/linguistic violation, for the semantic features of the adjective 'sour' do not fit with the semantic features associated with 'trains'. When 'sour' is predicated onto 'trains' an internal linguistic violation occurs relating to the knowledge about how words work, as opposed to the violation of world knowledge elicited by 'white', which does not possess this semantic violation.

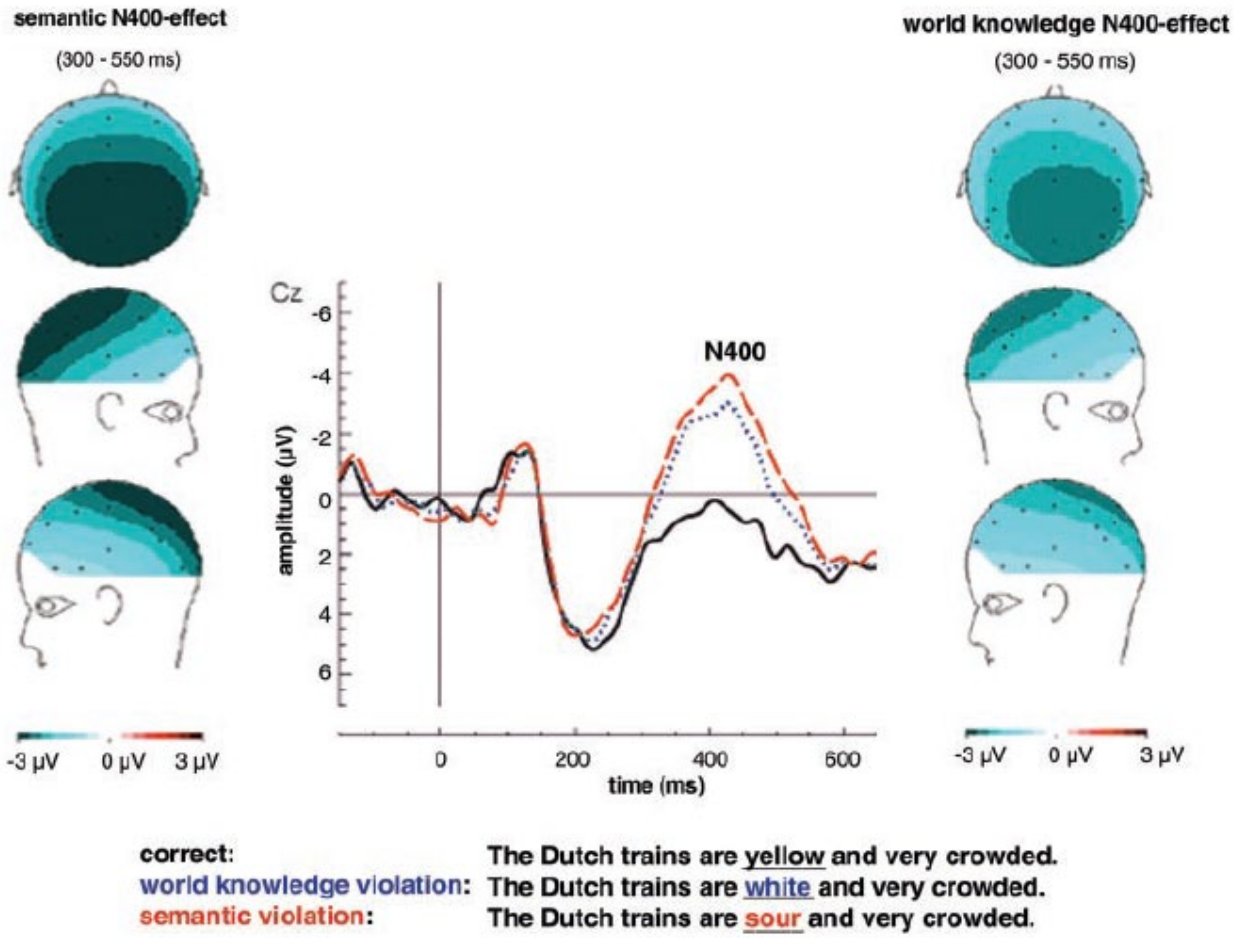

Figure 6: The grand average ERP for the Cz electrode, taken from Hagoort et al. (2004). Three sentence types/conditions: the correct condition is represented by the black line, the world knowledge violation condition is represented by the blue dotted line and semantic violation condition is represented by the red dashed line. ERPs are time locked to the presentation of the critical adjectives 'yellow', 'white' and 'sour'.

As results in Figure 6 show, both the world knowledge violation condition and the semantic violation condition elicited a significantly larger N400 response in comparison to the correct/true sentences (i.e. the control condition). Furthermore, there was a small, but significant, difference between the amplitudes of the world knowledge violation and semantic violation conditions, the latter being larger. However, the peak latency and topographic distributions of the two violation conditions were found to be the same. On the basis of this result, Hagoort et al concluded that world knowledge violations and semantic violations are comprehended in a similar fashion, processed in parallel via the same physiological process, a result consistent with a one-step, immediacy theory of comprehension and understanding.

Further supporting evidence that there is an immediate integration of lexical meaning and previously established knowledge, came from Nieuwland and Van Berkum (2006). In this study, they had their participants listen to short stories, such as that quoted below (and in Figure 7), in which a usually inanimate object is animated through a particular discourse narrative, giving rise to an interesting experimental prediction (detailed below). First, consider the following story, with the two N400 target phrases highlighted in bold:

> 'A woman saw a dancing peanut who had a big smile on his face. The peanut was singing about a girl he had just met. And judging from the song, the peanut was totally crazy about her. The woman thought it was really cute to see the peanut singing and dancing like that. The peanut was **salted / in love**, and by the sound of it, this was definitely mutual. He was seeing a little almond.'

Of the two predicates, 'salted' and 'in love', the former is the more likely one to occur in ordinary conversation, for the predicate is inanimate and so too (usually) is the use of the term 'peanut'. However, the peanut in this short story has been injected with a level of animation, from the dialogue building up to the two predication options. Thus, there is a tension between linguistic knowledge (i.e. the inanimate interpretation: 'the peanut is salted') and a more general discourse and world knowledge (i.e. the animate interpretation: 'the peanut is in love'), which gives rise to a concrete N400 predication, that if world knowledge is integrated in the same comprehension mechanism, then the inanimate predicates 'was salted' should elicit and increased N400 relative to the animate predicate 'was in love'. A result of this kind would be consistent with the Hagoort et al. (2004) experiment, discussed above, and other earlier results in this area (Hagoort & van Berkum, 2007; van Berkum et al., 2003).

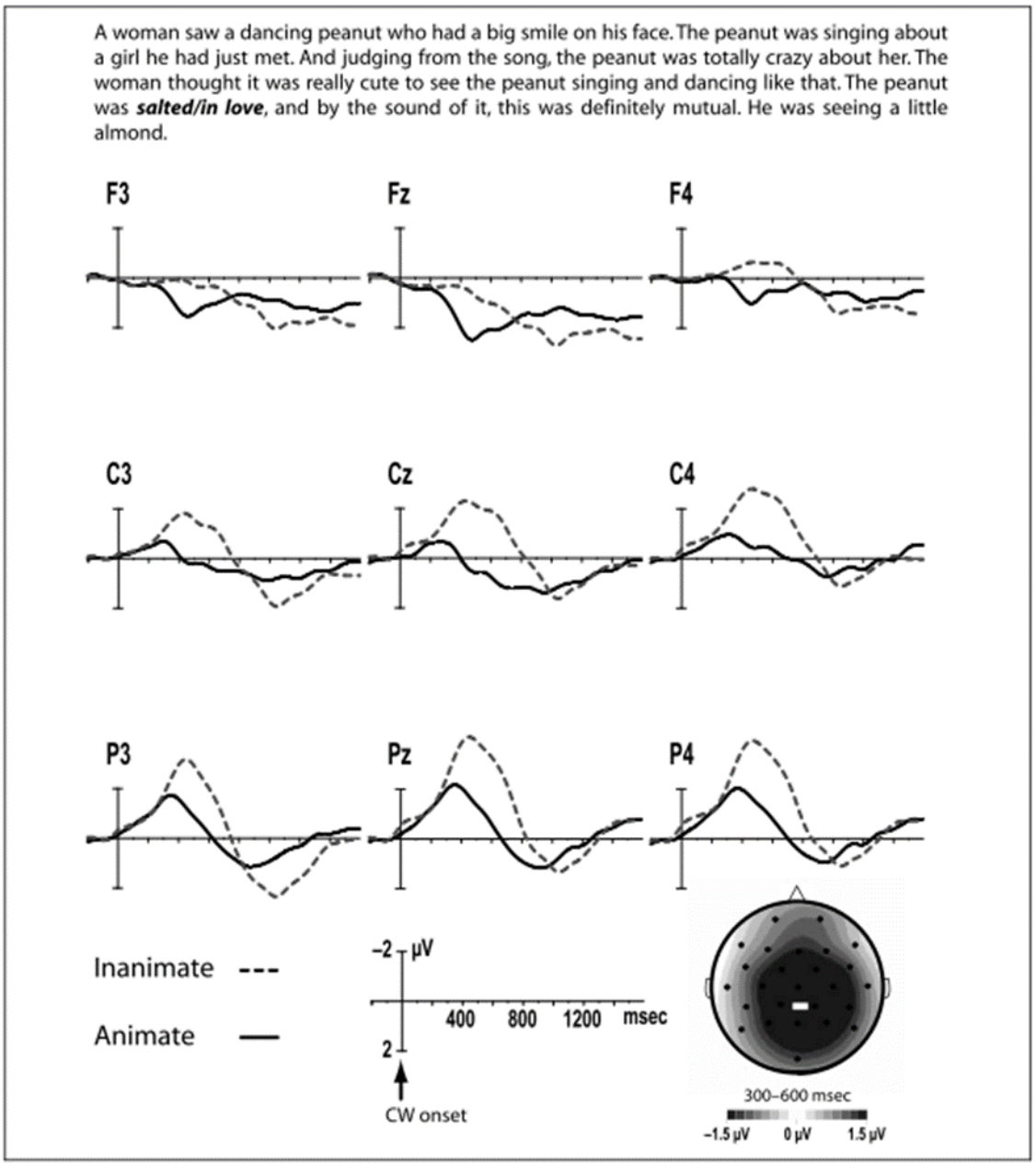

Figure 7: Grand average ERP waveforms of the N400 effects triggered by the correct, but contextually disfavoured, predicate (salted), in comparison to an incorrect, but contextually favoured, predicate (in love) (Nieuwland & Van Berkum, 2006)

As the results of Figure 7 show, the inanimate predicate (was salted), does indeed correlate with an increased N400 relative to the animate predicate (was in love). This result can be understood to support the view that prior context is an integral aspect of language comprehension, an inherent part of a singular, one-step mechanism that balances the integration of typical word meaning with the broader context of previous discourse, which is part of our world knowledge.

Another important EEG study, with respect to N400 and world knowledge, was conducted by Metzler et al. (2014) in which they examine world knowledge violations in the context of the personal knowledge of the participants; the participants being 16 schizophrenia patients and 16 neurotypical controls.

In this study, they visually presented 3-word-long German sentences, such as 'I am happy' and 'she is anxious'; the former they called the self-reference (SR) condition, the latter the other-reference (OR) condition. The SR condition required the participants to evaluate whether the adjective was true of themselves, the OR condition whether the adjective was true of a predefined friend or family member, with the appropriateness of the adjective indicated with a yes/no button press; 'yes' equivalent to 'congruent', 'no' equivalent to 'incongruent'. This study was highly influential in the design of the N400 components in the empirical investigations presented in §5.6 and §5.7.

The EEG figure presented in this paper is not clear enough to read (2014, pp. 536, Fig. 3). Nevertheless, their results showed an increased N400 effect during the processing of incongruent sentences, in comparison to congruent sentences, but this only occurred in the neurotypical controls and not the schizophrenia patients. This means that neurotypical controls were sensitive to violations of their own world knowledge, consistent with the above-mentioned literature, but that patients with schizophrenia were not sensitive to this type of violation, thus their neuronal behaviour in the comprehension of language is not statistically normal, more on this point in §5.7 and 5.10.

5.4 The P600 ERP

Another important language ERP is the P600. The P600 is a positive waveform that begins at approximately 500ms post-stimulus onset with a peak in amplitude around 600ms, with a broadly posterior distribution. The P600 has been demonstrated to index syntactic processing and is modulated by a variety of types of syntactic violations and syntactic complexities: (Ainsworth-Darnell, Shulman, & Boland, 1998; Coulson, King, & Kutas, 1998; Hagoort & Brown, 1994; Hagoort, Brown, & Groothusen, 1993; Hagoort, Brown, & Osterhout, 1999; Osterhout, 1997; Osterhout & Holcomb, 1992; Osterhout, Holcomb, & Swinney, 1994).

An early picture of the P600 in language was documented by Hagoort et al. (1993), in which they compared ERPs where there was a violation of agreement between a finite verb and subject NP in an ungrammatical condition, which was not present in a grammatical condition. Hagoort et al. (1993) presented sentences (in Dutch) such as 1 and 2 below, where the word

that renders the sentence ungrammatical or grammatical is bolded. Sentences below 1 and 2 represent the English translation:

1. Het verwende kind **gooit** het speelgoed od de grond
   The spoiled child **throws** the toys on the floor

2. *Het verwende kind **gooien** het speelgoed op de grond
   *The spoiled child **throw** the toys on the floor

The typical pattern of results elicited for the grammatical and ungrammatical condition are shown in Figure 8 , below.

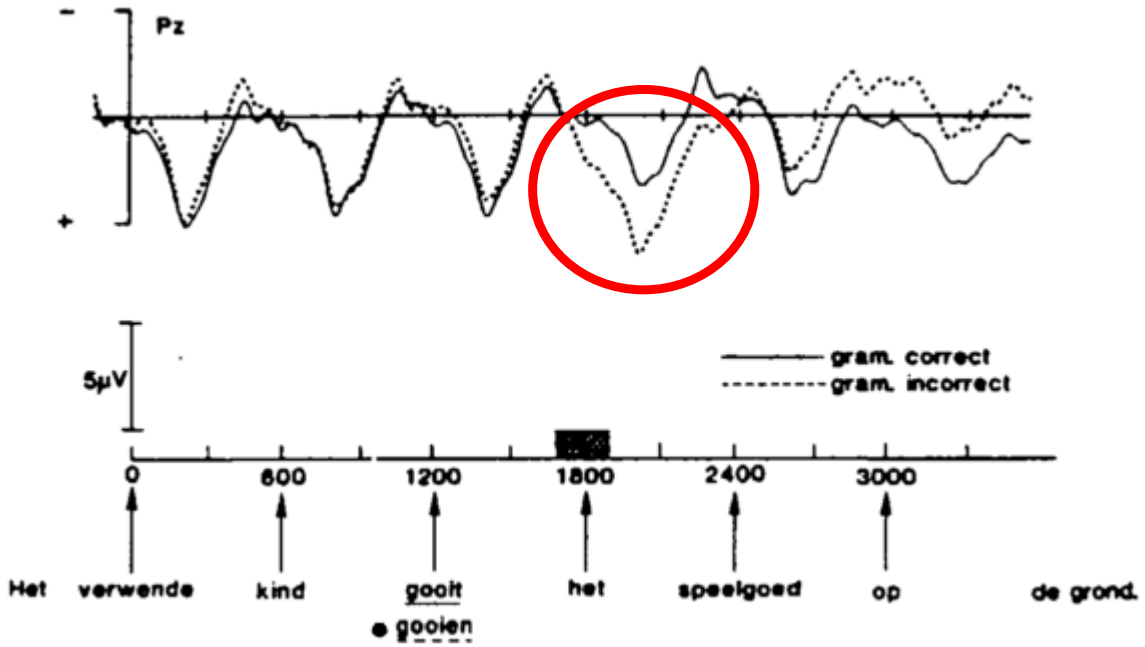

Figure 8: The grand average ERP waveform for the Pz electrode, taken from Hagoort et al. (1993). The grammatically incorrect sentence correlates with an increased P600 effect ≈600ms post critical word (CW) onset (i.e. the word that makes sentence ungrammatical: 'gooien') in comparison to the P600 associated with the grammatically correct sentence: 'goolt'.

A similar result was found with phrase structure violations. In Dutch, the obligatory word order is adjective-adverb-noun, Hagoort et al. (1993) violated this by changing the order of the adjective and adverb:

3. De echtgenoot schrikt van de nogal emotionele **reactie** van zijn vrouw
   The husband is startled by the rather emotional **reaction** of his wife

4. *De echtgenoot schrikt van de emotionele nogal **reactie** van zijn vrouw
   *The husband is startled by the emotional rather **reaction** of his wife

The results of these phrase structure violations – shown in Figure 9, below – also reveal an increased P600 component for the ungrammatical, phrase structure violation relative to its grammatical counterpart.

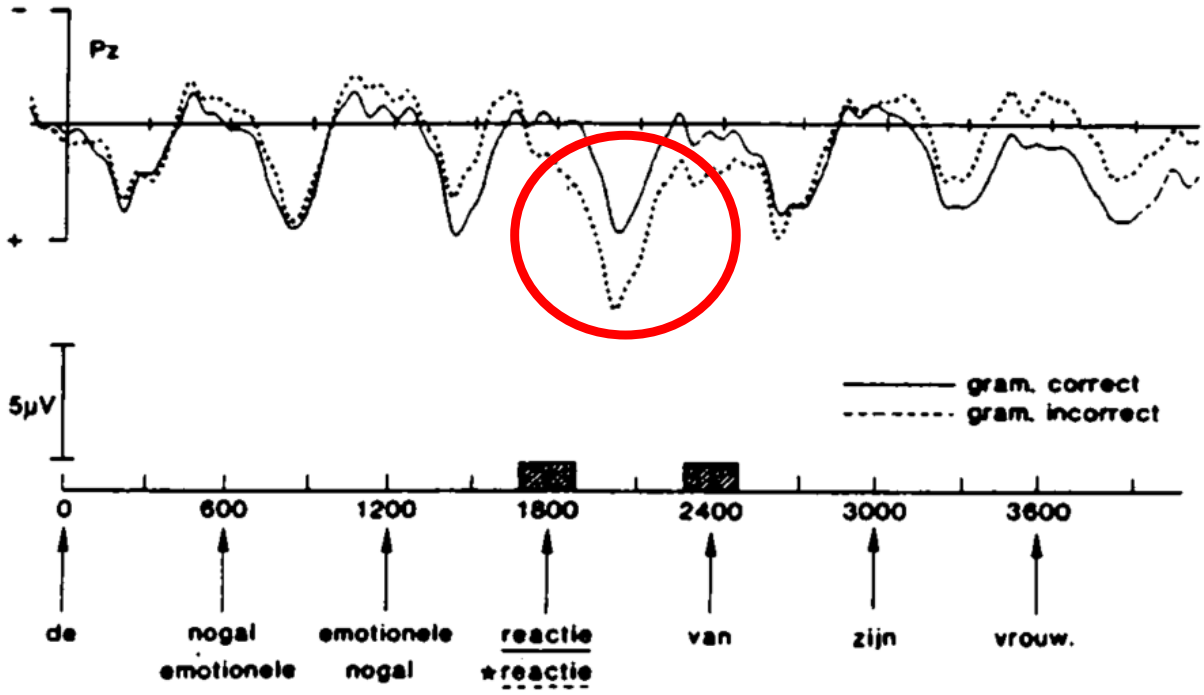

Figure 9: The grand average ERP waveform for the Pz electrode, taken from Hagoort et al. (1993). Similar setup to that described in Figure 8, only with phrase structure, grammatical violations.

The P600, conceived as a measure and index of syntactic processing, synchronises neatly with the Vosse and Kempen (2000) unification model (u-model) of parsing, and the predictions that arise from this theory, discussed in §2.5 and §3.4. However, one caveat to this link is that it is also possible that the P600 is consistent with many other parsing models, but, as demonstrated in previous chapters, this model also makes predictions consistent with many other theoretical and empirical findings, as well as arguments presented in this thesis (see Chapter 3).

To recap, the u-model argues that each word is associated with a grammatical, structural frame. The structural frames associated with these words are elicited when these words are individually comprehended, and as each word unfolds in a sentence, so too does the assembly and binding of the associated chunks of structure (see Figure 2, §2.5). Importantly, the binding of structure is not permitted when a root node of some particular piece of structure does not find space to occupy the foot node of incoming structure or when agreement features are violated (e.g. number, gender, tense, case, person, word order checks). In a similar sense, the P600 is modulated by syntactic violations and complexity. The suggested connection here is that the P600 indexes the cognitive binding of the structural frames postulated in the u-model. Specifically, that there is a positive correlation between the P600 amplitude and the difficulty of a particular binding operation, so that the more difficult it is to bind two structural frames, between two words, the greater the P600 amplitude. This theory explains why we see an increased P600 when there is a syntactic violation in a sentence, for the structural frames cannot easily combine. The use of 'increased' in the previous sentence means relative to a sentence without a syntactic violation, which by comparison, is easier to bind.

The relation between the Vosse and Kempen u-model and the P600 was first suggested by Hagoort (2003a) and has since been supported by several neurophysiological findings, some of which were outlined in §2.5. Importantly, in the context of ERP research, is the Hagoort, Wassenaar, and Brown (2003) study with aphasic patients. In this study, Hagoort et al demonstrated that agrammatic aphasics performed poorly on behavioural/syntactic comprehension tests, in comparison to neurotypical controls and nonagrammatic aphasics. Moreover, the agrammatic aphasics, unlike the nonagrammatical patients and neurotypical controls, failed to show the typical P600 response to grammatical violations, where there was an agreement violation between the subject and the verb of a sentence (e.g. 'The girls pay the baker and **takes** the bread home'). As is shown in Figure 10 (below), the agrammatic patients were neurophysiologically insensitive to this grammatical violation. Based on this finding, the proposition was that the cortical areas lesioned in these patients (left perisylvian areas) are necessary for typical parsing. On this point, Hagoort (2003a) argues that this area, especially the left superior temporal cortex, is responsible for the retrieval of the syntactic frames stored in association with individual words, whilst the left inferior frontal cortex is responsible for their binding into larger structures.

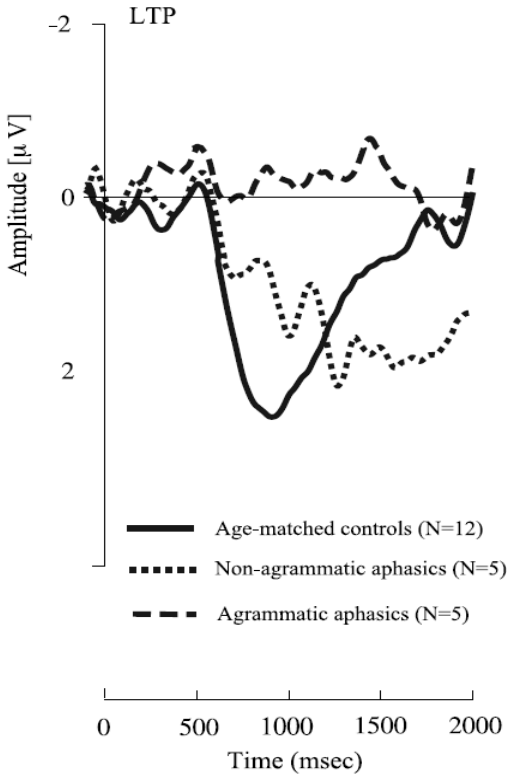

Figure 10: ERP waveform at temporal electrode site, in which agrammatic aphasics do not show the characteristic P600 response to a grammatical violation, in comparison to Broca's aphasics and neurotypical controls. Taken from Hagoort et al. (2003).

The first indication that the N400 and P600 could be dissociated from one another – that the P600 is independent of some possibly confounding factors relating to meaning – came from Hagoort and Brown (1994), where they demonstrated that the P600 also occurs in sentences where the usual semantic requirements have been entirely removed:

1. The boiled watering-can **smokes** the telephone in the cat.
2. *The boiled watering-can **smoke** the telephone in the cat.

The comparative results of these sentence types – where 1 is semantically incongruent, but grammatical, and 2 is semantically incongruent (in the same way as 1), but is ungrammatical – reveals a P600 increase in the ungrammatical condition (*smoke) in comparison to the grammatical condition (smokes). Even though no semantically congruent meaning was communicated, the language system was able to parse the sentence using grammatical information (see Hagoort & Brown, 1994, p. 70, Fig3.8, for visualisation of result) or (see

Hagoort et al., 1999, pp. 287, Fig9.3, for English interpretation). More recent EEG research, using frequency analysis, supports this dissociation between syntactic and semantic processing, with Beta frequency activation (15-30Hz) representing syntactic processing and Gamma frequency activation (30-90Hz) representing semantic processing; violations of which, result in attenuation of these frequencies in comparison to syntactically/semantically correct sentences (Bastiaansen & Hagoort, 2015; Hald, Bastiaansen, & Hagoort, 2006; Lam et al., 2016).

5.5 The interplay between the N400 and P600 ERPs

As discussed in the previous two paragraphs, these semantic and syntactic ERPs, the N400 and P600 respectively, are taken to be separate processes in the brain, but exactly how these components work together to generate comprehension is still not clear. Hagoort (2003b) and others: (Ainsworth-Darnell et al., 1998; Friederici, Steinhauer, & Frisch, 1999; Gunter, Stowe, & Mulder, 1997; Osterhout & Nicol, 1999) have attempted to elucidate the nature of this interplay by combining syntactic and semantic violations, in ERP experiments, on the same critical word (CW). In combining these violations on the same CW, both an increased N400 and P600 is expected to arise, therefore their interaction can be examined.

For example, Hagoort (2003b) examined simultaneous semantic and syntactic violations in Dutch, within an NP. Specifically, the syntactic violations involved a mismatch in gender and number features between the determiner and noun. In Dutch, nouns have one of two grammatical genders: common gender or neuter gender. When part of an NP, the gender of a noun is marked by the determiner: ‘de’ is used for common gender nouns and ‘het’ is used for neuter gender nouns. Furthermore, in number agreement: ‘de’ has to be used for definite plural NPs and ‘het’ with a singular NP. It was in the mismatch of the determiners and nouns, in gender and number features, that served as syntactic violation condition of the experiment.

The semantic violations involved an anomalous combination of adjective and noun (e.g. honest umbrella). As stated above, both syntactic and semantic violations were presented in the same NP, with the noun serving as the CW (see Table 4 for an example of the stimuli and conditions).

Sentence	Condition
De kapotte *paraplu* staat in de garage	Control
Het kapotte *paraplu* staat in de garage	Syntactic violation
De **eerlijke** *paraplu* staat in de garage	Semantic violation
Het eerlijke *paraplu* staat in de garage	Syntactic and Semantic violation
English translation: The$_{com}$/**The**$_{neut}$ broken/**honest** *umbrella*$_{com}$ is in the garage	

Table 4: The critical word (i.e. the noun) is italicised in red font, the incorrect determiner (Het) and anomalous adjective (eerlijke) is in bold, representing the syntactic and semantic violations, respectively. In the English translation, 'com' is an abbreviation for article/noun of a common gender and 'neut' for neuter gender article.

Hagoort found that the N400 was increased with the addition of the syntactic violation. Whereas, the P600 was not affected by the addition of a violation of meaning, yet the syntactic violation conditions still resulted in a significantly larger P600 effect (see Figure 11 for visual representation of grand average waveforms).

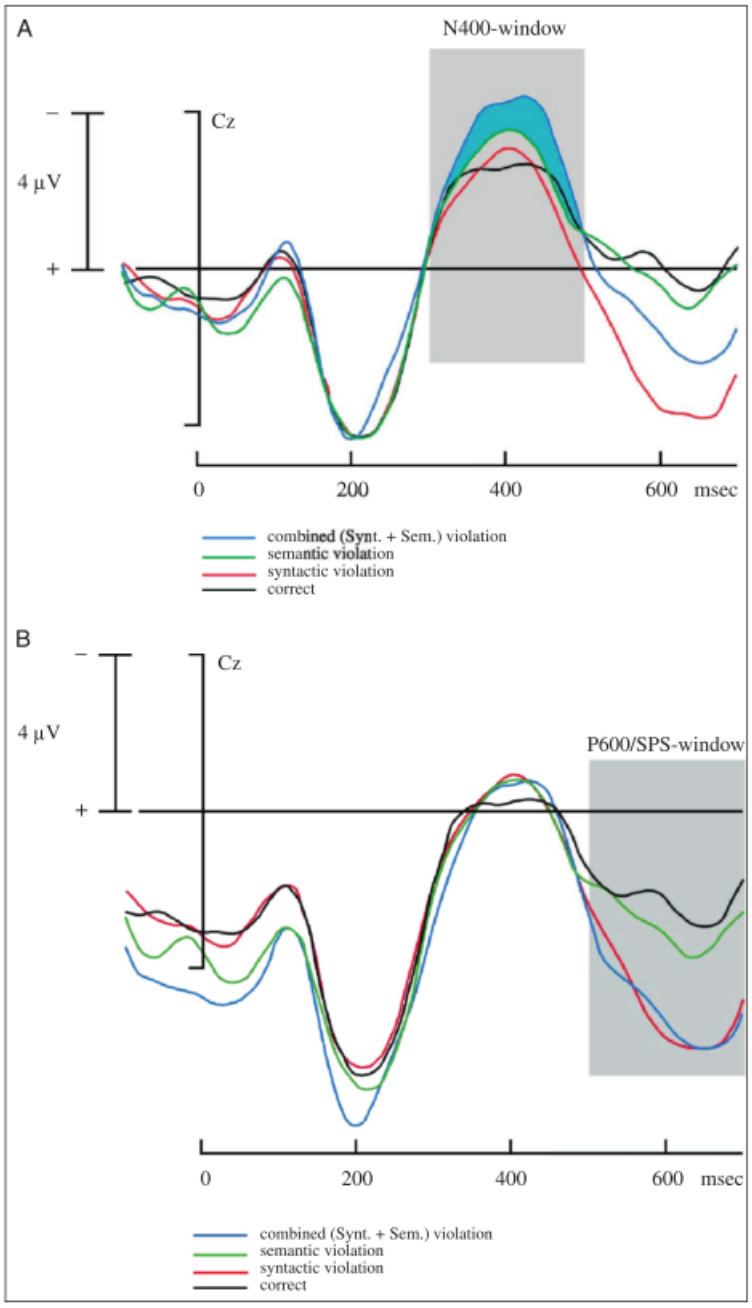

Figure 11: Grand average ERP waveforms for the four conditions types examined in Hagoort (2003b): Correct (black line), syntactic violation (red line), semantic violation (green line), combined syntactic/semantic violation (blue line). Figure A represents the N400, Figure B the P600; with adjusted baseline (further discussed in §5.6.2.4).

To fully understand the N400 and P600 components and their interactions, it is important to examine them in the differing forms of language use, including the world knowledge perspective discussed in §5.3. This ability, to produce and comprehend true and false utterances, is one of the defining and most useful characteristics of human communication. It

is where language connects with an individual speaker /listener's knowledge of the world. The absence of this perspective is a limitation to the experiments discussed in this section, on the interplay between the N400 and P600, for they do not explore meaning, and meaning violations, from within this framework; especially not with respect to self-related information. An examination of these components and their interplay, in the world knowledge context, will serve to increase the ecological understanding on the physiological basis of language comprehension, for these types of comprehension can be said to be approaching real-life settings.

At this point in the thesis, with the above knowledge gap identified, the tact changes from theoretical to experimental. Hereafter, two EEG experiments are presented, in which, the N400 and P600 ERPs are empirically investigated in the context of self-related, world knowledge, examined as individual and combined ERPs. The first experiment is a comprehensive experiment with neurotypical adults. The second experiment is a pilot experiment with patients with a diagnosis of schizophrenia, a psychiatric disorder typically characterised as involving an altered form of selfhood; more on this in §5.7.1.

With the data in hand from these experiments – examined and embedded within the context of the previous literature – consideration can then be given as to what neurophysiological evidence, as a representation of typical comprehension behaviour, can be factored into the PLTb profile of self. Furthermore, this evidence may also be integrated into the separate and ever-growing literature on language comprehension physiology. The extent of the above-mentioned integrations will depend on the robustness of results, which, in neuroscience, as in most experimental sciences, will require further experimentation and replication to verify its truth, or to confirm/avoid its falsification (Popper, 1963, 1980).

5.6 EEG experiment in neurotypical adults

5.6.1 Introduction

In the present EEG experiment, with neurotypical adults, the individual N400 and P600 components, and their interplay, were investigated in the context of world knowledge and syntactic violations, as an extension of the work described in §5.3, 5.4 and 5.5. Sentences were visually presented to participants, which they were tasked to read and comprehend whilst undergoing EEG recording. The sentences were constructed to allow four different conditions to arise: Gram+Agree, Gram+Disagree, Ungram+Agree and Ungram+Disagree. The

Gram/Ungram distinction related to whether the sentence was grammatical or ungrammatical, whereas the Agree/Disagree distinction related to whether the participant agreed or disagreed with what the sentence was saying. When the participant disagreed with what the sentence was saying, this was classified as a meaning violation in the context of world knowledge. The task required the participants to judge the appropriateness of trait adjectives with reference to themselves and a 3rd person male of their selection (see §5.6.2.2 for full details of task).

The design of the task was partly based on the work performed by Esslen, Metzler, Pascual-Marqui, and Jancke (2008) and Metzler et al. (2014) discussed above (§5.3), in which trait adjectives were judged during EEG recording. However, these experiments do not consider grammatical violations and neither do they consider the interaction between grammatical violations and world knowledge violations.

On the basis of previous research discussed in §5.3, 5.4 and 5.5, particularly that presented by Hagoort et al. (2004), Hagoort (2003b) and Nieuwland and Van Berkum (2006), the predictions of this study were as follows: Firstly, that there would be an increased N400 response for both of the Disagree conditions, in comparison to the Agree conditions (especially so for the grammatical conditions, based on Metzler et al. (2014)), as the former are considered to be semantic violations (of world knowledge). Furthermore, that the Ungram+Disagree condition would demonstrate an additional increase in N400 amplitude, in comparison to Gram+Disagree condition, boosted by the syntactic violation, a result consistent with Hagoort (2003b). Secondly, that there would be a significantly increased P600 amplitude for both of the Ungram conditions, in comparison to the Gram conditions (when appropriately adjusting for the N400 effect, see §5.6.2.4), but that there would be no difference between the two Ungram conditions, in that, the semantic violation brought about by the disagree decision, would not affect the P600 amplitude.

5.6.2 Methods

5.6.2.1 Participants

26 neurotypical adults (11 male, 15 female), aged between 18 and 69 years: (mean (M) ± standard deviation (SD): 39.7 ± 13.5 years), years in education: (M ± SD: 16.0 ± 3.2 years), gave written informed consent to participate in this experiment. All participants were free of significant past/present physical illness and were not using psychoactive drugs. Participants were screened to exclude significant current, past or family history of psychiatric illness. Participants had normal (or corrected normal) vision and hearing and were right-handed, as

assessed by the Edinburgh Handedness Inventory (Oldfield, 1971); see §5.7.2.2 for further details on this measure. Participants were monolingual, native English speakers, recruited via the online volunteer system at the Institute of Neuroscience, Newcastle University. Participants were compensated for their participation with a small financial honorarium (£20). All procedures were ethically approved by the Institute of Neuroscience, Newcastle University. The experiment was funded by two personally awarded grants: £1500 from the *North East Mental Health Foundation* and £2000 from the *Royal College of Psychiatrists*.

5.6.2.2 Experimental procedure and stimulus selection

Subjects were shown 120 (3-word-long) English sentences, word by word, in the centre of a computer screen. After the third word of the sentence, subjects were required to make a decision as to whether they thought the sentence was true (i.e. they agreed with the content of the sentence and thus was congruent with their world knowledge) or whether they thought it was false (i.e. they disagreed with the content of the sentence and so constituted a world knowledge violation), indicating this agree/disagree decision with a button press.

The sentences consisted of a pronoun, followed by a copular verb or conjunction, followed by an adjective, with the adjective functioning as the CW of the sentence, for it is only with the onset of the adjective where the possibility of truth and falsity arises. Half of the sentences (n=60) were grammatical and half were ungrammatical. This difference in grammaticality was also marked by the onset of the adjective, set up by the choice of preceding word (copular or conjunction, described below). Importantly, this design allowed for interactions between syntactic and world knowledge violations, for they both occurred with the onset of the adjective.

The overall content of the sentences related to whether a person (referred to by the pronoun), has a certain personality trait (specified by the adjective), or not, for example, one grammatical sentence was ‘I am happy’. Half of the sentences (n=60) began with the pronoun ‘I’ and the other half with the pronoun ‘He’. For the sentences beginning with ‘I’, the participant had to decide whether they thought the sentence was true or false with respect to their own personality, indicating this with a button press. For the sentences beginning with ‘He’, the participant had to decide whether they thought the sentence was true or false with respect to a third-person male. This third-person male was selected offline, prior to the experiment. The participant was asked to pick a male person that they knew well, in order that they can make sound judgments about their personality characteristics. The person that the

participant selected, would be the reference for the pronoun ‘He’, for the entirety of the experiment.

The second word of each sentence was always one of the following: ‘am’, ‘is’ ‘or’ ‘and’. When the copular verbs ‘am’ and ‘is’ were presented after the pronoun, they ensured that the overall sentence would be grammatical, confirmed as such with the subsequent onset of the adjective (e.g. ‘I am happy’, ‘He is sad’). These types of sentences formed the grammatical condition of this experiment. In contrast to this, when the conjunctions ‘and’ and ‘or’ were presented after the pronoun, they ensured that the overall sentence would be ungrammatical, also confirmed as such with the subsequent onset of the adjective (e.g. ‘I and dramatic’, ‘He or selfish’). These types of sentences formed the ungrammatical condition of this experiment; two conjunctions were used, instead of one, to prevent neural entrainment to the syntactic violations. Participants were required to make the agree/disagree decision in both grammatical and ungrammatical conditions. This was a straightforward task for the grammatical condition, for the participant simply comprehended the sentence as it was presented to them (the presentation process is illustrated in Figure 12, below), and then made a decision as to whether the sentence was true or false. The ungrammatical condition was more complicated, for in this condition, the conjunction (‘and’ or ‘or’) made the sentence ungrammatical when the adjective was presented, and thus the sentence was not straightforwardly comprehensible. Nevertheless, the participant was asked to read each word of the sentence as it unfolded and then continue to make the same agree/disagree decision, but doing so on the basis of the pronoun and adjective, only. In other words, they would read each word of the ungrammatical sentence (e.g. ‘I and dramatic’), but then they would make their agree/disagree decision based on the pronoun ‘I’ and the adjective ‘dramatic’. Therefore, in this example, the participant is the reference and they have to decide whether or not they are dramatic, and indicate this with a button press. This procedure is the same for the ‘He’ sentences.

In the ungrammatical condition, it was hypothesised that an ERP epoched around the CW adjective would shed light on three things. Firstly, the cognitive reaction to the syntactic violation; secondly, the cognitive, sentence-repair process; and thirdly, depending on the participant decision, the possible interaction between a world-knowledge meaning violation and a syntactic violation.

The third word of each sentence was a trait-related adjective (e.g. ‘happy’, ‘sad’, ‘dramatic’, ‘jealous’). 240 adjectives were used in total to counterbalance the participant

exposure to the adjectives across the different conditions (see Table 5 for overview of conditions). The adjectives and their associated frequency-of-use statistics were taken from the British National Corpus. This allowed adjective frequency-of-use to be counterbalanced across conditions, ensuring that the different amount of uses of adjectives in language were not an experimental confound. Adjectives deemed especially emotive were not selected.

Finally, to elicit an approximately equal amount of agree and disagree decisions from the participants, a small behavioural experiment was performed prior to this EEG experiment. In this behavioural experiment, 10 participants (different from the EEG participants) were given the full list of adjectives and were asked to indicate whether they would agree or disagree with the adjectives with respect to their own personalities. A score for each adjective was computed based on the amount of agrees/disagrees elicited, across the participants. Based on these scores, the adjectives were parsed out across the experimental lists to try and ensure a sufficient amount of agrees and disagrees were elicited (see Table 6 in §5.6.3 for frequency of agrees/disagrees elicited in this experiment).

Sentence Types	**Condition**	**Number of presentations**
I am ADJECTIVE	Grammatical	30
His is ADJECTIVE	Grammatical	30
		Grammatical, N = 60
I or ADJECTIVE	Ungrammatical	15
He or ADJECTIVE	Ungrammatical	15
I and ADJECTIVE	Ungrammatical	15
He and ADJECTIVE	Ungrammatical	15
		Ungrammatical, N = 60

Table 5: Breakdown of sentence types in this experimental design.

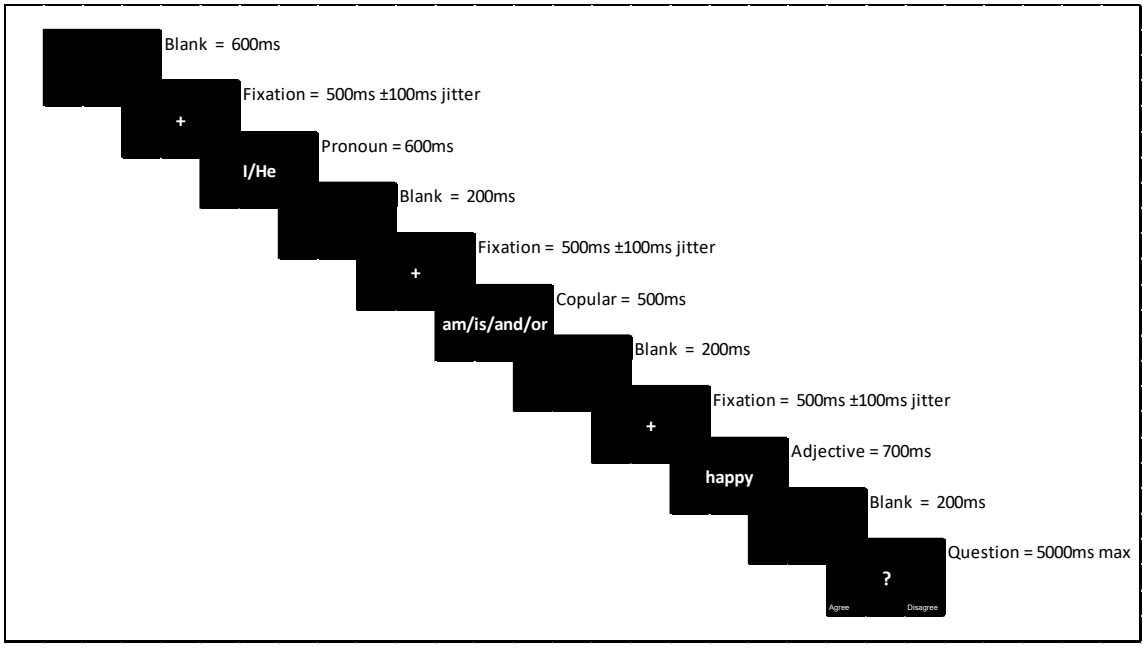

Figure 12: Each word was presented in the centre of the EEG lab computer screen. The individual stimuli unfolded as pictured and described above. Fixation points were presented in between words to reduce ocular artefacts and were subject to random variation in the timing offset (±100ms) to de-correlate the training of neurophysiological responses to the repetitive presentation of similar types of stimuli. After the 700ms CW adjective presentation and a 200ms blank screen, a question mark appeared in the centre of the screen prompting the agree or disagree decision. These decisions were indicated with a button press via the two Ctrl buttons on a standard QWERTY keyboard. 'Agree' and 'Disagree' were presented in the bottom left and bottom right of the screen, approximately in line with the two Ctrl buttons to remind the participants of the button allocation. The first 13 participants indicated Agree with the left Ctrl button and Disagree with the right Ctrl button. This was reversed for the subsequent participants, to control for possible motor effects associated with hand dominance and button pressing.

5.6.2.3 Data collection

EEG data was recorded using a Synamps 2 amplifier: Neuroscan, USA. 32 silver/silver chloride electrodes were placed on the scalp, embedded in an elasticated cap: Easy Caps, Germany. Electrodes were setup in accordance with the 10-20 system (Klem, Luders, Jasper, & Elger, 1999). The AFz electrode served as the ground reference electrode, and the right mastoid as reference electrode, which was re-referenced (offline, post-experiment) to an average of the linked mastoids (see Figure 19, §8.1 for overview of electrode placement). VEOG were

recorded between electrodes placed on the nazion and infraorbital ridges below the centre of each eye. HEOG were recorded between electrodes placed on the outer canthus of both eyes. All electrode impedances were kept below 10kΩ. ERP's were initially computed from 200ms prior to adjective onset (the initial baseline period, see §5.6.2.4 for details on baseline selections) to 1000ms after adjective onset. Standardised analysis procedures, in the Neuroscan Scan4.4 software, were used on the EEG data. Initially, the EEG data was visually inspected for gross artefacts, which were manually removed. This was followed by a blink-correction procedure (Semlitsch, Anderer, Schuster, & Presslich, 1986) to remove the blink EOG artefact. Recordings were filtered with a band pass: low pass filtering set at 30Hz (6db/oct), high pass filtering set at 0.5Hz (24db), sampled at a rate of 500Hz. All channels, apart for VEOG, with voltage deflection ±75μV, were excluded, with an overall average trial loss from artefact rejection ≈31%.

5.6.2.4 Statistical Analysis

The N400 and the P600 are reported as maximally detectable in central-parietal areas (Duncan et al., 2009; Kutas & Federmeier, 2011). Therefore, the CPz electrode was *a priori* selected as the electrode of interest. Other central-parietal electrodes are visually presented, but are not statistically analysed (see figures in §8.2). CW onset is defined as time zero (0ms). The N400 was quantified as the mean amplitude of individual participant data in the 300-500ms time window, the P600 as the 500-700ms time window. These windows were selected *a priori* on the basis of previous literature (Hagoort, 2003b). The mean amplitudes for the N400 were computed relative to a 200ms baseline preceding the CW onset. The P600 was recalculated in two different manners, to explore and compensate for the preceding N400 differences. As discussed by Hagoort (2003b, p. 892) one way to compensate for the baseline problem, from adjacent ERPs, is to compute peak-to-peak amplitudes from the N400 to the P600 (Coles, Gratton, Kramer, & Miller, 1986). However, determining the peak values in individual participants' average waveforms, in the absence of a sharply defined peaks, especially in the P600 component, is prone to error. Consequently, Hagoort used the individual mean N400 amplitudes, for each condition, as the baseline for the P600 analysis. This was the first form of P600 analysis in this experiment.

In the second analysis, the P600 was computed using the individual mean amplitudes between 480-500ms as the baseline. The reason for this second analysis was to compensate for a possible flaw in Hagoort's approach. Namely, that in utilising the entirety of the N400

window, as a proxy baseline for the peak to peak amplitude, implies that all data points in this window, affect those elicited in the P600 window, and thus, should be corrected for statistically. However, there seems no reason to suppose that this level of influence is the case, especially in the early, pre-peak, stages of the N400 window. Both types of analysis were performed to explore whether any differences would arise in the results.

5.6.3 Results

Condition	**Mean number of trials included**
Gram+Agree	M = 25.5, SD = 6.5
Gram+Disagree	M = 25.7, SD = 5.9
Ungram+Agree	M = 26.5, SD = 7.2
Ungram+Disagree	M = 24.4, SD = 6.3

Table 6: The mean (M) and standard deviation (SD) of the number of trials included, per participant, across conditions, in the statistical analysis of this experiment.

The CPz mean amplitude values were entered into an omnibus repeated-measures ANOVA for the N400 and the P600 ERPs separately with Grammaticality [2 levels: Gram, Ungram] and Decision [2 levels: Agree, Disagree] as within-subject factors. Mauchly's test of sphericity was not computed, since the variables analysed have only two levels. *Post hoc* analysis was performed where there were reliable main/interaction effects. This involved paired sample t-tests examining permutations of the 4 conditions: Gram+Agree, Gram+Disagree, Ungram+Agree and Ungram+Disagree. The Bonferroni correction was applied as the confidence interval adjustment. Partial ETA Squared (η^2_p) effect sizes (small = 0.02, medium = 0.13, large = 0.26) and Cohen's *d* effect sizes (small = 0.2, medium = 0.5, large = 0.8) were reported for the repeated-measures ANOVAs and paired sample t-tests, respectively (see Lakens, 2013; for overiew of effect sizes).

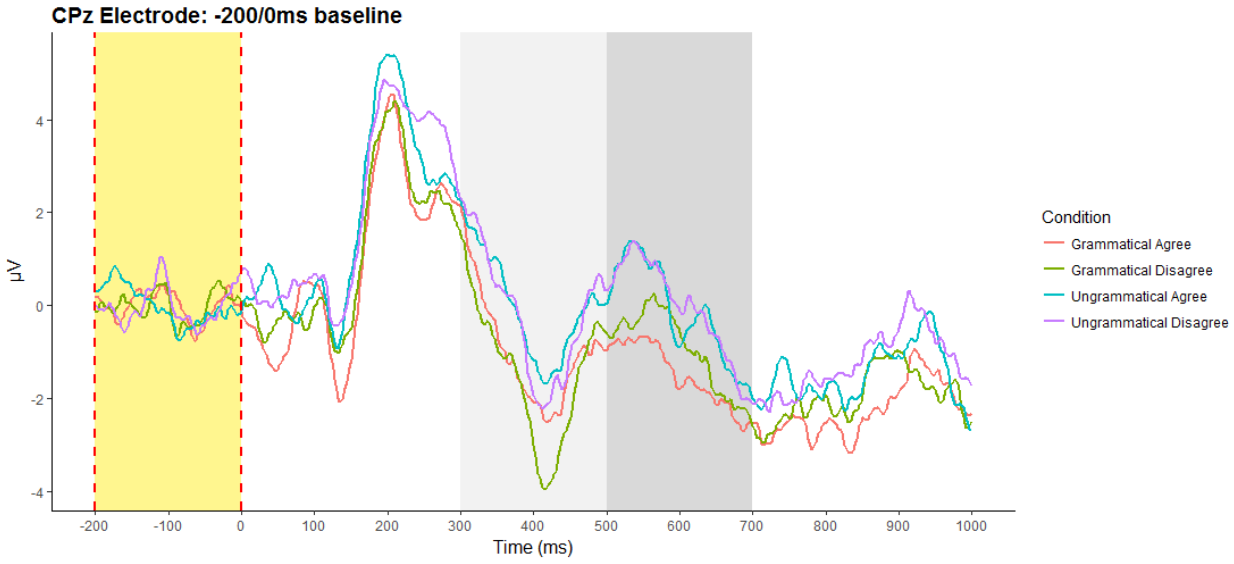

Figure 13: Grand average ERP waveforms from the CPz electrode for the four experimental conditions, with the CW presented at 0ms. The khaki area represents the baseline (-200ms-0ms), the light grey area represents the N400 window (300-500ms) and the dark grey area represents the P600 window (500-700ms).

5.6.3.1 **N400**

Repeated measures ANOVA examining the N400 resulted in a significant main effect of Grammaticality, $F(1,25) = 10.081$, $p = 0.004$, $\eta^2_p = 0.287$. The grammatical conditions (Gram+Agree and Gram+Disagree) elicited an increased amount of negativity in the N400 window (M ± SE: -0.683μV ± 0.429μV) in comparison to the ungrammatical conditions (Ungram+Agree and Ungram+Disagree) (M ± SE: 0.681μV ± 0.500μV), see Figure 13 above and Figure 14 below; the latter is zoomed-in on the N400 window. There was no main effect of Decision, $F(1,25) = 1.590$, $p = 0.219$, $\eta^2_p = 0.060$, nor a Decision * Grammaticality interaction effect $F(1,25) = 0.064$, $p = 0.803$, $\eta^2_p = 0.003$.

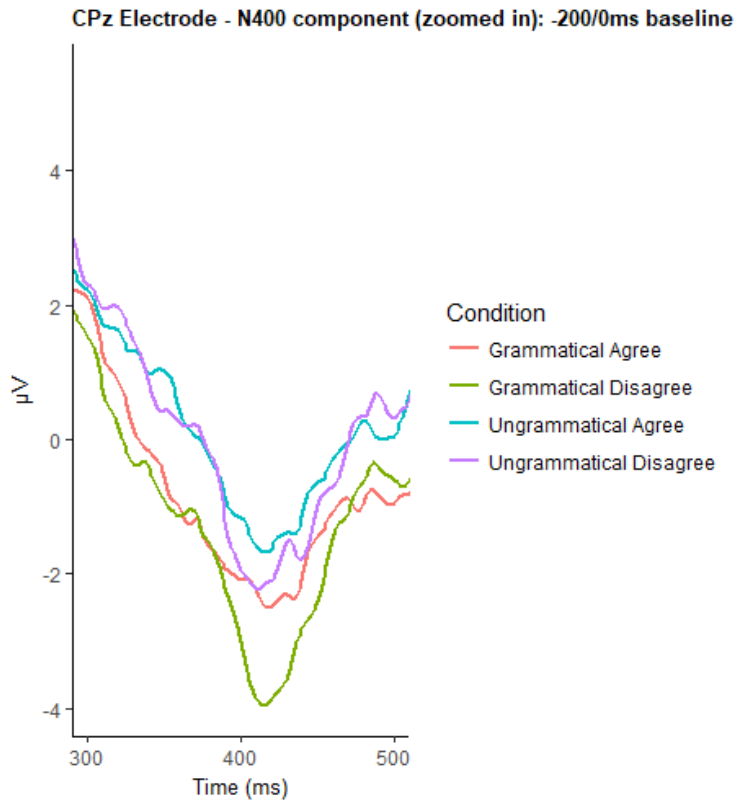

Figure 14: Grand average ERP waveforms from the CPz electrode for the four experimental conditions. The waveform is zoomed-in on the N400 window (300ms-500ms), baseline corrected by the -200ms-0ms window.

To further explore the significant main effect of Grammaticality and the *a priori* experimental hypotheses (§5.6.1), this ANOVA was followed-up by 5 paired samples t-tests (two-tailed); 4 to explore the effect of Grammaticality and 1 to explore the world knowledge violation between Gram+Disagree and Gram+Agree. There was a significant difference between Gram+Agree (M ± SE: -0.436μV ± 0.484μV) and Ungram+Agree (M ± SE: 0.847μV ± 0.511μV); t(25) = -2.355, p = 0.027, d = 0.462 – a non-significant difference between Gram+Agree and Ungram+Disagree (M ± SE: 0.515μV ± 0.597μV); t(25) = -1.476, p = 0.152, d = 0.289 – a significant difference between Gram+Disagree (M ± SE: -0.929μV ± 0.439μV) and Ungram+Agree; t(25) = -4.333, p < 0.001, d = 0.850 – a significant difference between Gram+Disagree and Ungram+Disagree; t(25) = -2.763, p = 0.011, d = 0.542 – and a non-significant difference between Gram+Disagree and Gram+Agree; t(25) = 1.157, P = 0.258, d = 0.426.

5.6.3.2 P600

As discussed in §5.6.2.4, the P600 was computed relative to two different baselines (300-500ms and 480-500ms) to compensate for the preceding N400 differences. A repeated measures ANOVA examining the P600 with the 300-500ms baseline (see Figure 15, below) revealed a significant main effect of Decision, $F(1,25) = 11.712$, $p = 0.002$, $\eta^2_p = 0.319$, with the disagree conditions (Gram+Disagree and Ungram+Disagree) (M ± SE: 0.129μV ± 0.370μV) eliciting an increased amount of positivity in the P600 window in comparison to the Agree decisions (Gram+Agree and Ungram+Agree) (M ± SE: -0.642μV ± 0.129μV). There was no main effect of Grammaticality, $F(1,25) = 0.562$, $p = 0.460$, $\eta^2_p = 0.022$, nor a Decision * Grammaticality interaction effect $F(1,25) = 1.648$, $p = 0.211$, $\eta^2_p = 0.062$.

To further explore this significant main effect of Decision, this ANOVA was followed by 4 paired-samples t-tests (two-tailed). There was a significant difference between Gram+Agree (M ± SE: -0.897μV ± 0.537μV) and Gram+Disagree (M ± SE: 0.187μV ± 0.336μV); $t(25) = -3.274$, $p = 0.003$, $d = 0.642$ – a significant difference between Gram+Agree and Ungram+Disagree (M ± SE: 0.071μV ± 0.458μV); $t(25) = -2.394$, $p = 0.025$, $d = 0.470$ – a significant difference between Gram+Disagree and Ungram+Agree (M ± SE: -0.386μV ± 0.450μV); $t(25) = 2.079$, $p = 0.048$, $d = 0.408$ – and a non-significant difference between Ungram+Agree and Ungram+Disagree; $t(25) = -1.371$, $p = 0.183$, $d = 0.269$.

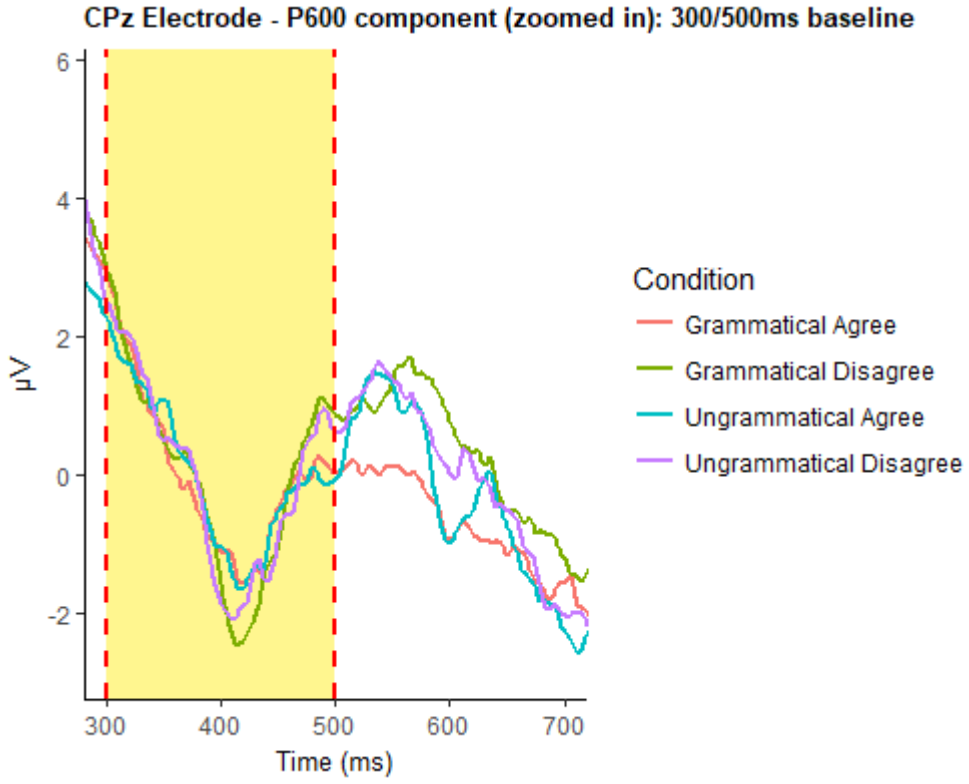

Figure 15: Grand average ERP waveforms from the CPz electrode for the four experimental conditions. The waveform is zoomed-in on P600 window (500ms-700ms), with the N400 window (300ms-500ms, the khaki area) functioning as the baseline.

For the second computation, a repeated measures ANOVA examining the P600 with the 480-500ms baseline (Figure 16, below) revealed no significant main or interaction effects. Therefore, follow-up paired sample t-tests were not performed.

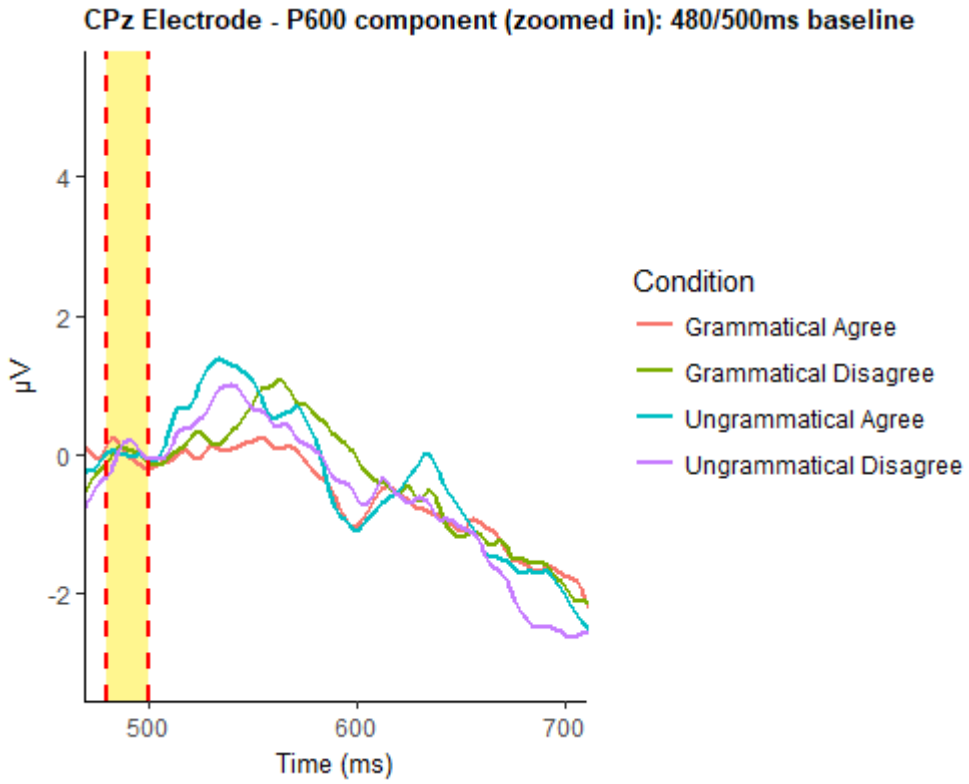

Figure 16: Grand average ERP waveforms from the CPz electrode for the four experimental conditions. The waveform is zoomed-in on P600 window (500ms-700ms), with a 480ms-500ms window (the khaki area) functioning as the baseline.

5.6.4 Discussion

The results of this experiment suggest that in the comprehension of a written sentence, a violation in grammar correlates with a significant reduction in N400 amplitude on the CW of the sentence, in comparison to a grammatical sentence. However, on the basis of the *post hoc* t-test analysis of the N400 component, in conjunction with a visual inspection of N400 grand average waveform (Figure 14), it could be argued that this difference in grammaticality is

mostly driven by the increased amplitude of the Gram+Disagree condition (the world knowledge violation condition), and not the Gram+Agree condition, which has a more comparable amplitude to the Ungram conditions; especially Ungram+Disagree (non-significant difference). Furthermore, on the basis of previous literature, it was predicted that the Disagree conditions (especially the grammatical) would elicit an increased N400 relative to the Agree conditions, because the former was a semantic violation and semantic violations reliably elicit a relative increase in N400 (see §5.3). However, in opposition to this interpretation, it could be highlighted that the Gram+Agree condition was significantly different from Ungram+Agree and was tending towards a significant difference between Ungram+Disagree ($p = 0.15$, two-tailed), and was not significantly different from Gram+Disagree. Therefore, the proposal that a violation in grammar does correlate with a reduction in N400 seems reasonable, statistically speaking. In view of this conflict, three possible conclusions are proposed, either: (1) the differences in grammaticality correlate with differences in N400 response, and that grammatical sentences correlate with larger N400 amplitudes in comparison to ungrammatical sentences, or (2) the Gram+Disagree condition, perhaps as an instance of a world-knowledge violation, drove the main effect of Grammaticality, therefore this result should be interpreted as consistent with previous world knowledge violation findings, particularly those presented by Metzler et al. (2014), or (3) a combination of both effects (1-2) were at play in this experiment. Based on the *post hoc* p-values, conclusion (1) appears to be the most probable conclusion, but more research is required to further delineate between these possibilities.

With respect to the P600, the results of this experiment are not conclusive, for they are shown to vary depending on the choice of baseline. Prior to statistical analysis, a visual inspection of the P600 grand average waveform with the -200ms to 0ms baseline correction (Figure 13), suggests that the ungrammatical conditions elicit an increased P600 amplitude relative to the grammatical conditions. A result of this kind would be consistent with previous literature on the P600 (see §5.4) and potentially in line with the N400 conclusion (1), explained above (e.g. in the ungrammatical condition, the reduced N400 is followed by an increased P600, the latter representing a repair mechanism of the former). However, this view does not take into consideration the differing waveform behaviours prior to the P600 window, therefore does not necessarily capture the relative change in voltage – in this instance, the peak-to-peak amplitudes from the N400 to the P600 – which is the measurement of interest in this experiment. To correct for this, the P600 was calculated in two different manners (see §5.6.2.4

for full reasoning behind this choice). Firstly, it was computed relative to a 200ms baseline preceding the P600 window (i.e. the baseline was the N400 window, 300-500ms) and secondly, the P600 was computed relative to 480-500ms baseline, a smaller window directly preceding the P600 window. The results of the first analysis revealed an unexpected difference between Decision, with the Disagree conditions eliciting an increased P600 amplitude relative to the Agree conditions. The *post hoc* t-test analysis revealed that this main effect of Decision was mainly driven by the differences between Gram+Agree and Gram+Disagree, since the differences between Ungram+Agree and Ungram+Disagree were non-significant. However, the results of the second analysis revealed no significant differences whatsoever. Consequently, it is difficult to make any concrete conclusions on the P600 ERP in this experiment, with respect to its role in language comprehension. That said, this experimental design does highlight methodological considerations for researchers interested in examining the interplay of the N400 and P600. Specifically, that the choice of baseline can affect the results. This point on baselines also applies to other research that is investigating interactions between ERP components, and is especially relevant to those ERPs that are directly adjacent to each other, in time. There is more discussion on circumventing this baseline issue in §5.8 and 5.9.

To conclude, this study builds upon previous ERP findings that examine semantic processing (in the context of world knowledge violations) and syntactic processing (in the context of phrase structure violations). The main findings are as follows: (1) the N400 ERP is modulated by the grammaticality of a sentence and that violations in grammar correlate with a significant reduction in N400 amplitude, relative to a grammatical sentence. This conclusion should be interpreted with some caution, for reasons outlined in the first paragraph of this section. And (2), that caution is required when designing and analysing the interplay between ERPs (here, the N400 and P600) for the selection of baseline can radically alter the results.

The full limitations of this experiment and the following pilot experiment are outlined in §5.8.

5.7 EEG pilot experiment in schizophrenia patients

5.7.1 Introduction

Schizophrenia is a severe and often chronic mental disease, defined by the possession of one or more of the following abnormalities: ‘delusions, hallucinations, disorganized thinking (speech), grossly disorganized or abnormal motor behaviour (including catatonia), and

negative symptoms' (American Psychiatric Association, 2013, p. 87). Importantly, for this thesis, schizophrenia is often linked to a loss of self or attenuated selfhood (Kircher & David, 2003; Sass & Parnas, 2003), and the theories of the self in schizophrenia are as diverse and essentialist in content as those developed on the self, detailed in Chapter 1. Consequently, these theories face the exact same criticisms, outlined in §1.6.

The following is a pilot EEG experiment with patients with a diagnosis of schizophrenia (hereafter, 'SZ patients') and neurotypical controls (hereafter, 'controls'). One of the aims of this pilot is to generate data on how language comprehension processes unfold in SZ patients, in the context of world knowledge and syntactic violations, utilising the same experimental paradigm as the previous experiment, §5.6. However, because this is a pilot, the sample size is much smaller than the previous experiment; therefore all results are preliminary.

Other more concrete aims of this pilot involve the assessment of the experimental design: to see whether the task is acceptable to the patients and would lead to valid results; to establish realistic parameters for recruitment, and importantly, to allow the primary researcher (William Jones) to acquire experience working with patients, including: establishing informed consent, administering sensitive psychological measures, identifying the barriers and facilitators to successfully employing neuroimaging techniques with patients, and developing an awareness of the regulatory issues, NHS structures and policies required to perform clinical research (see Thabane et al., 2010 for further details and rationale behind pilot studies). This information could potential inform the development of follow-up, full-scale, patient EEG projects.

Given the N400 results from the first experiment, in conjunction with the ambiguity identified with the P600 results, and the preliminary nature of this analysis, the focus in this pilot is on the N400 ERP, only. Several studies have demonstrated abnormal N400 processing in schizophrenia, typically showing reduced N400 responses in SZ patients in comparison to controls (Jackson et al., 2014; Kiang, Kutas, Light, & Braff, 2007; Kumar & Debruille, 2004; Metzler et al., 2014; Mohammad & DeLisi, 2013; Onitsuka, Oribe, Nakamura, & Kanba, 2013; Salisbury, 2008; Sitnikova, Salisbury, Kuperberg, & Holcomb, 2002; Wang, Cheung, Gong, & Chan, 2011). A common explanation for these results is that SZ patients possess an overactive semantic memory network, coupled with a verbal working memory deficit (see Salisbury, 2010). However, it is not completely clear that an overactive semantic memory network makes clear predictions for the N400, for an increased N400 (in SZ patients, relative to controls) could be equally consistent with this theory. The working memory aspect of this theory will be briefly

explored below, see §5.7.2.5 for details. In addition to this, the role of verbal IQ will also be examined, since this has been shown to correlate with linguistic aspects of schizophrenia (Çokal et al., 2018).

Based on this previous research, it is expected that SZ patients will elicit a reduced mean N400 amplitude, in comparison to controls. Further fine-grained predictions of the individual conditions are not here being made, since the reliability of the results will likely be compromised by the small sample size. It is also expected that the controls will replicate the main effect of Grammaticality from the first experiment.

5.7.2 Methods

5.7.2.1 Participants

In this pilot experiment, there was 22 participants in total: 11 SZ patients and 11 controls. The SZ patient inclusion criteria included: (1) pre-existing diagnosis of schizophrenia, in line with DSM-5, (2) aged between 18 and 60 years old, (3) English as their first and only language (i.e. monolingual, native speakers of English), (4) right-hand-dominant, assessed by the Edinburgh Handedness Inventory (see §5.7.2.2 for details), and (5) able to provide written informed consent. Control participant inclusion criteria included numbers (2-5), above. SZ patient exclusion criteria included: (6) a pre-existing primary diagnosis of alcoholism or substance dependence, (7) NART Verbal IQ score < 85 (see §5.7.2.2), (8) organic disease of the brain including significant head injury, stroke, tumour or epilepsy, and (9) severe dyslexia. Control participant exclusion criteria included numbers (6-9) above, but in addition to these: (10) is a first degree relative of a family member with schizophrenia and (11) has a personal history of mental illness or substance abuse.

SZ patients were recruited via psychiatrists working in the Newcastle upon Tyne Hospitals Trust, Campus for Ageing and Vitality (CAV), Wolfson Research Centre. William Jones, the author of this thesis, was trained and supervised by Professor Hamish McAllister-Williams and Professor Douglas Turkington to assess the above criteria and to administer the cognitive assessments described below in §5.7.2.2. The controls were recruited via the Voice Volunteer Service and the online volunteer system at the Institute of Neuroscience, Newcastle University.

Participants were compensated for their participation with a small financial honorarium (£40), and all travelling to and from the CAV was booked and paid for in advance

by William Jones, supported by the aforementioned grants (see §5.6.2.1). All procedures of this study were ethically approved by local Research Ethics Committee (REC reference: 16/SC/0424) for the Newcastle upon Tyne Hospitals Trust, in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki. All aspects of the ethical application were developed by William Jones.

5.7.2.2 Participant assessments

This pilot experiment was divided into two separate sessions, both occurring within one week of each other. The first session was used to establish capacity and consent, and to assess the participants on a range of cognitive measures. The second session was dedicated to the EEG testing. The first session involved the employment of the following measures: the Edinburgh Handedness Inventory (EHI) (Oldfield, 1971), the National Adult Reading Test (NART) (Nelson, 1982), the working memory digit span test (Humstone, 1919) and the Brief Psychiatric Rating Scale (BPRS) (Overall & Gorham, 1962). Years in education was also recorded.

The EHI is a measurement scale used to assess a person's hand dominance, a standard measure for imaging studies examining language processes. The NART is a measure for estimating premorbid ability (or verbal IQ) and has been shown to strongly correlate with the full WASI-IV IQ score (Bright, Hale, Gooch, Myhill, & van der Linde, 2018; Wechsler, 2008). The NART involves the pronunciation of 50 words that become increasingly irregular. The score from the NART is then converted into IQ via the following formula: NART predicted WAIS-IV FSIQ = −0.9775 × NART error + 126.41 taken from (Bright et al., 2018, p. 7). The working memory digit span test measures short-term, verbal working memory (Woods et al., 2011). Here, the participant is required to recall the order of a digit sequence in forward and reverse order. The sum of these scores gives a total digit span for the participant. Finally, the BPRS is a measure used to assess the positive, negative and affective symptoms associated with psychosis. In total, 18 symptoms are assessed, with scores on each symptom rated from 1 (not present) to 7 (extremely severe).

5.7.2.3 Experimental procedure and stimulus selection

As per §5.6.2.2.

5.7.2.4 Data collection

As per §5.6.2.3.

5.7.2.5 Statistical analysis

As per §5.6.2.4. However, the focus here is on the N400 ERP. Furthermore, in this analysis, an ANCOVA is performed to explore the variance in amplitude differences when verbal IQ and working memory capacity are controlled for.

5.7.3 Results

The SZ patients and controls were matched on sample size, gender, age, handedness (approximately), and all participants had normal (or corrected normal) vision and hearing. SZ patients and controls were significantly different in years of education, working memory digit span, NART score, and the BPRS. See Table 7 below for more details on these measures.

Measure	**Controls**	**SZ patients**	**P value**
n of participants	11	11	-
Gender	Male = 7, Female = 4	Male = 7, Female = 4	-
Handedness (EHI)	Right = 11, Left = 0	Right = 10, Left = 1	-
Age	M = 44.58, SD = 15.55	M = 45.17, SD = 10.11	p = 0.910
Years in education	M = 17.21, SD = 2.62	M = 13.83, SD = 2.25	p = 0.003
Digit span score	M = 16.67, SD = 5.10	M = 11.75, SD = 3.65	p = 0.013
NART score	M = 113.70, SD = 5.04	M = 102.87. SD = 11.01	p = 0.007
BPRS score	M = 1.08, SD = 0.12	M = 2.13, SD = 1.11	p < 0.001

Table 7: Summary of participant demographics and scores on cognitive measures. The P value column represents the p value difference between the two groups, derived from two-tailed independent samples t-tests.

The CPz mean amplitude values were entered into a two-way mixed factorial repeated measures ANOVA for the N400 ERP with Grammaticality [2 levels: Gram, Ungram] and Decision [2 levels: Agree, Disagree] as within-subject factors and Group [SZ patients, Controls] as the between-subjects factor. Mauchly's test of sphericity was not computed, since the variables analysed have only two levels. Levene's test indicated that the variances were homogenous for all levels of the repeated measure variables; all significance values > 0.05.

Post hoc analysis was performed on reliable main/interaction effects and where specific predictions were made, see §5.7.1. The Bonferroni correction was applied and effect sizes were reported.

Repeated measures ANOVA examining the N400 revealed no significant within-subject main effects of Grammaticality, $F(1, 20) = 1.747$, $p = 0.201$, $\eta^2_p = 0.080$; Decision, $F(1,20) = 1.644$, $p = 0.214$, $\eta^2_p = 0.076$, or interaction between these, $F(1,20) = 0.064$, $p = 0.803$, $\eta^2_p = 0.003$. There was a significant main effect of Group, $F(1,20) = 5.416$, $p = 0.031$, $\eta^2_p = 0.213$, with Controls (M ± SE: -1.119μV ± 0.486μV) eliciting a significantly increased N400 response in comparison to SZ patients (M ± SE: 0.481μV ± 0.486μV), see Figure 17 and Figure 18 (below) for visual representation of this difference; the decision to combine conditions in these figures was made to make them clear enough to interpret. There was no Group by within-subject factor interaction for either Grammaticality or Decision, or three-way interaction.

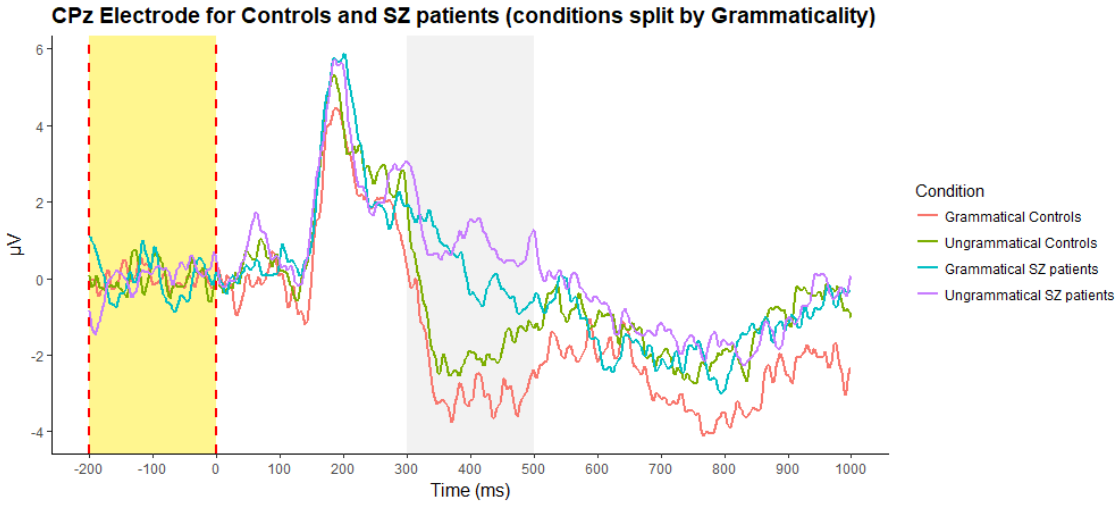

Figure 17: Grand average ERP waveforms from the CPz electrode for the two experimental groups – Controls and SZ patients – with the four conditions split by Grammaticality and Group. CW presented at 0ms. The khaki area represents the baseline (-200ms-0ms), the light grey area represents the N400 window (300-500ms).

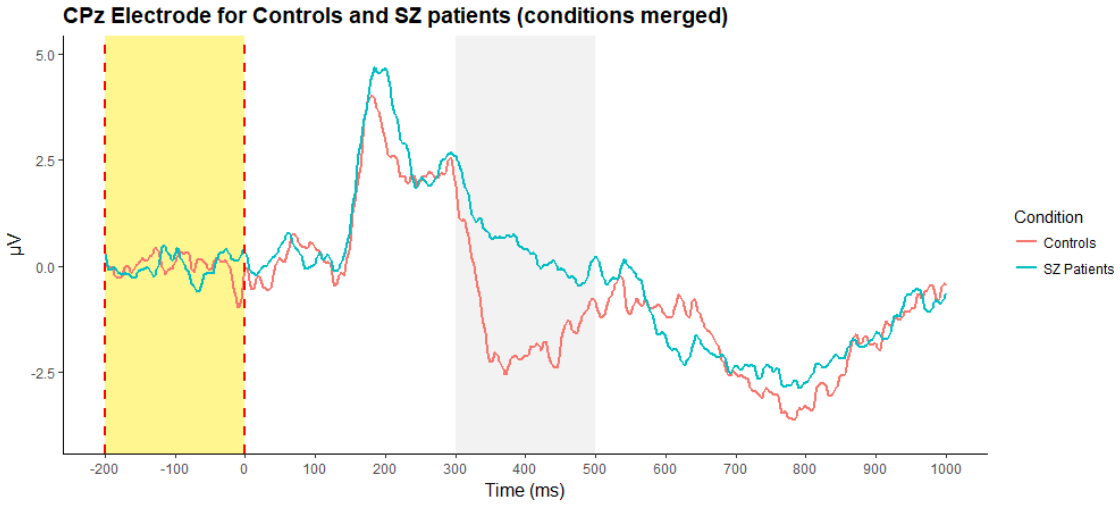

Figure 18: Grand average ERP waveforms from the CPz electrode for the two experimental groups – Controls and SZ patients – with the four conditions merged (averaged together) into one waveform.

A follow up repeated measures ANOVA, examining the N400 in SZ patients, resulted in a non-significant main effect of Grammaticality, $F(1,10) = 0.227$, $p = 0.644$, $\eta^2_p = 0.022$, and a non-significant effect of Decision, $F(1,10) = 1.126$, $p = 0.314$, $\eta^2_p = 0.101$, and a non-significant interaction effect of Grammaticality * Decision, $F(1,10) = 0.322$, $p = 0.583$, $\eta^2_p = 0.031$.

Follow up repeated measures ANOVA examining the N400 in controls resulted in a non-significant main effect of Grammaticality, $F(1,10) = 2.540$, $p = 0.142$, $\eta^2_p = 0.203$, and non-significant effect of Decision, $F(1,10) = 0.519$, $p = 0.488$, $\eta^2_p = 0.049$, and a non-significant interaction effect of Grammaticality * Decision, $F(1,10) = 0.120$, $p = 0.736$, $\eta^2_p = 0.012$.

An ANCOVA was conducted to explore the differences in Group whilst controlling for verbal IQ and working memory capacity, using the participant's NART and digit span score as measures of these, respectively. The Shapiro-Wilk test of normality, Levene's test for equality of variances and the homogeneity of regression slopes assumption, all indicated significance values > 0.05. ANCOVA results showed a non-significant effect of Group when verbal IQ and working memory capacity were controlled for, $F(1,18) = 1.452$, $p = 0.224$, $\eta^2_p = 0.075$ (a reduction in η^2_p of 0.138, compared to Group without the covariates), with estimated adjusted means of: controls (M ± SE: -0.865μV ± 0.574μV) and SZ patients (M ± SE: 0.226μV ± 0.574μV).

5.7.4 Discussion

The results of this experiment suggest that in the comprehension of a written sentence, patients with a diagnosis of schizophrenia (SZ patients) process meaning differently than neurotypical controls (controls), consistent with previous literature on the topic (see §5.7.1). The processing of meaning is here indexed by changes in N400 amplitude, and in this experiment, SZ patients had a significantly reduced N400 response in comparison to Controls, where all four conditions (Gram+Agree, Gram+Disagree, Ungram+Agree and Ungram+Disagree) were averaged together (see Figure 17 and Figure 18).

The significant within-subject effect of Grammaticality reported in the first EEG experiment, was not replicated in this experiment, in either the controls ($F = 2.540$, $p = 0.142$) or SZ patients ($F = 0.227$, $p = 0.664$). A power calculation performed in G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) utilising the effect size derived from the first EEG experiment, $d = 0.506$, assuming an alpha of 0.05 (two-tailed), indicated that 53 participants would be required for 95% power and 33 participants for 80% power. A power calculation utilising the effect size derived from the controls of the second (pilot) EEG experiment, $d = 0.368$, alpha 0.05 (two-tailed), indicated that 97 participants would be required for 95% power and 59 participants would be required for 80% power. A power calculation of the Grammaticality * Group interaction, utilising the effect size derived from the second (pilot) EEG experiment, $d = 0.123$, alpha 0.05 (two-tailed), 2 groups, 2 repetitions, indicated that 218 participants (total sample size) would be required for 95% power and 132 participants for 80% power. Considering these numbers, it is possible that the failure to replicate results from the first experiment may be due to a lack of power; more on this in future directions (§5.9), below.

To further explore the significant Group difference, working memory capacity and verbal IQ were added to the analysis as covariates. This addition resulted in a non-significant result of Group, suggesting that a proportion of the difference between controls and SZ patients can be explained by a reduced working memory capacity and verbal IQ in the SZ patients.

5.8 Study limitations

There were four internally/personally identified limitations (L1, L2, L3, L4) in the experimental paradigm employed in the above experiments, and a range of observations/limitations noted from the pilot experiment. Firstly (L1), with respect to experimental design, the way the N400 ERP, as an index of meaning comprehension, is examined in the ungrammatical conditions is potentially questionable. To recap, in half of the

trials of this experiment, participants had to make an agree/disagree decision when the sentences were ungrammatical (e.g. 'I and dramatic' and 'He or selfish'). Consequently, the sentences in the ungrammatical condition were not straightforwardly comprehensible, arguably requiring cognitive effort to repair their sentential/propositional meaning before the participants could make their agree/disagree judgements. This point is supported (anecdotally) by the feedback received from some of the participants, who reported a sense of jarring and a need for additional thinking when the CW was presented in the ungrammatical conditions. If this criticism holds, it follows that you cannot have a meaningful sentence, and a violation of that meaning, whilst the sentence is still in a state of ungrammaticality. Therefore, it seems impossible to have a truly simultaneous semantic and syntactic violation condition, for in order for the meaning to be violated, the sentence must first be meaningful, which requires that the sentence be grammatical. This proposition is consistent with the theoretical work developed on referential meaning in §2.4 and 3.3; that reference is a grammatical phenomenon. However, the experimental design does still provide an insight into the physiological reaction to a syntactic violation and/or the sentence-repair processes, which is then followed by a meaning decision (i.e. the design explores how the brain tries to circumvent ungrammaticality in creating meaning). This is exactly the type of ability that might be lacking in SZ patients (see Çokal et al., 2018). Furthermore, the stronger conclusions made on the N400 ERP – primarily those in the first experiment – are not affected by this aspect of the design, for this compared the differences between the grammatical and ungrammatical conditions, which the design adequately captures.

The second limitation (L2) relates to the relatively low, but still reasonable, number of trials per participant/condition in the first experiment (see Table 6). This point is in view of Woodman's (2010, p. 7) recommendation of between 30-60 recordings per participant/condition. This point is further compounded by the results from the effect size calculations presented in §5.7.4, which indicate that a large amount of participants are required to secure the conventional levels of adequate power (i.e. 80 and 95%). However, the results expressed with confidence in these experiments are those in which the four conditions are collapsed into two (e.g. the N400 Grammaticality effect in the first experiment and the Group effect in the second experiment), thus, the recommended numbers have been approximately satisfied, but still could be improved. More on this limitation in §5.9.

The third limitation (L3) relates to the pronouns 'I' and 'He' and the copular verbs and conjunctions 'am', 'is', 'and' and 'or'. The processing of these sentence components were not examined in this analysis.

The final limitation (L4) of the study relates to the analysis plan, namely, the *a priori* selection of the analysis windows (e.g. N400 = 300-500ms, P600 = 500-700ms). Arguably, it is good scientific practice to select these prior to examining the results, for it requires that a prediction be formulated which could be clearly false. However, this approach is lacking in flexibility and thus may not capture the ERPs in an ideal fashion. One possible way to overcome this issue, which does not involve an unscientific level of data wrangling/fishing, would be to utilise Independent Component Analysis (ICA) (Brown, Yamada, & Sejnowski, 2001; Groppe, Makeig, & Kutas, 2009; Makeig, Jung, Bell, Ghahremani, & Sejnowski, 1997). In simple terms, ICA is an automated solution to isolating the electrical activity of neurons, of finding independent sources of activity. Moreover, ICA can be performed without the need for baseline correction (Hanson, 2017, p. 40), so could potentially lessen the above-described problem with the P600 analysis (see §5.6.4).

One of the main aims of the pilot experiment was to assess the feasibility of the design for use with SZ patients (i.e. to see whether the task was acceptable to the patients and would lead to valid results). On this point, it was concluded that the experimental task was difficult for the patients to perform, especially the ungrammatical conditions (see §5.6.4 for further explanation of difficulty). Therefore, this aspect of the design was a potential confound to the results. Furthermore, recruitment of patients was challenging, partly because of the low prevalence of patients (see Kirkbride et al., 2012), but also because of the patient's reluctance to participate, driven by their high levels of paranoia. In short, it was concluded that to ensure that the N400 and P600 components be the examined in an efficient manner (i.e. in a possible, larger scale, patient study) the experimental design would benefit from certain alterations, detailed in the following section. These alterations aim to lessen the effects of the above-outlined limitations.

5.9 Future directions

Based on these two EEG experiments, a set of future directions have been identified. Firstly, ICA is a definite area to explore in future EEG/ERP research. As mentioned above, in §5.8, utilising ICA could help to capture ERP responses in a more valid and unbiased manner (L4), avoiding the baseline issue encountered in the first experiment.

Secondly, for reasons now outlined, in moving forward in this research area, it would be better to examine the N400 and P600 ERPs independent of one another, in separate, parsimonious and ecological experiments. This realisation is based on the insight that in statistical-based science 'simplicity is the ultimate sophistication' (Kass et al., 2016, p. 4), and that one should start with simple approaches and only add complexity as required. There would be several advantages to approach, in the context of the advancing the current experimental design. For example, in exploring the processing of world knowledge, it would have been sufficient to examine these in one condition only, specifically, the grammatical, self-reference, 'I' condition. In this scenario, the experiment would only present sentences of the type 'I am *adjective*' and the participants would assess these for truth and falsity. In utilising one condition to explore this phenomenon, there could be an increased number of trials per participant, from 30 (in the current design) to a possible 240 (the total number of adjectives garnered for this experiment). This alteration would lessen the impact of L2 above (i.e. low number of trials), as well as the outlined difficulty of patient recruitment, resulting in greater power to detect real physiological differences. Furthermore, a task with only grammatical 'I' sentences would have been easier for the participants to complete, which was also identified as an area of concern for the SZ patients. Note also that making this adjustment would likely not change the level of participant engagement in the experiment, for several participants reported that judging their own personality traits was a stimulating task. These points on simplification, through the isolation of conditions, also holds true for future designs examining the P600.

Consistent with the above tone, other future directions for EEG language research should consider including experimental designs that examine sentence level processing, for this involves an ecological examination of language use, beyond that present in the examination of single word processing. One method to examine sentence level comprehension is to utilise frequency oscillation analysis over the duration of a sentence, and to compare this with other sentence conditions (see Bastiaansen & Hagoort, 2015; Lam et al., 2016); see §5.2 for more details. In addition to this, a multi-model neuroimaging approach to language comprehension would be advantageous, as it can offset the disadvantages associated with individual neuroimaging paradigms (e.g. the poor spatial but good temporal resolution of EEG and the poor temporal but good spatial resolution of fMRI). As of September 1st 2018, I have been running a simultaneous function near-infrared spectroscopy (fNIRS) and EEG experiment, with neurotypical adults, which examines audibly presented grammatical and ungrammatical sentences (i.e. scrambled versions of the grammatical sentences). Statistical analysis will

examine the stimuli at the sentence level, and of particular interest is the relationship between the blood oxygenation level dependent (BOLD) signal in the left hemisphere language regions and the evoked EEG frequency oscillations, across conditions. This experiment was funded by a personally awarded grant: £12,534, from the *Newcastle upon Tyne Hospitals NHS Charity*.

5.10 EEG, the PLTb profile of self, and schizophrenia

Based on previous EEG literature and the evidence generated in the above experiments (but especially the former), one can say with relative confidence that the N400 and P600 ERPs index the comprehension of meaning and syntax, respectively; where meaning includes the world knowledge of the individual. If one accepts this proposition – and to accept this, one need not hold that these components only index language – then these can (potentially) be integrated into the PLTb profile of self. These components, represented as waveforms and measured statistically, can be considered as observable/triangulable behaviours; where behaviours are considered as a broad range of actions that are produced by individuals, organisms, objects and/or systems.

Integrating this behaviour into the PLTb self profile requires that one first establishes what constitutes the criterion of typical and non-typical behaviour, this way one will be able to make classification decisions. In Chapter 4, the normal distribution curve was advocated as a means of calculating the limits of this criterion, with within ±2 SDs of the mean (μ) constituting typical behaviour (i.e. mid95%) and more than ±2 SDs of the μ constituting non-typical behaviour (i.e. end5%). Applying this principle to N400 and P600 components means that a person is comprehending language typically (at least, neurologically speaking) if they produce waveforms within the mid95% range of a particular reference class (in this case, 'all humans' since we are exploring the self in humans) and non-typically if they produce waveforms within the end5% range.

One cannot say with absolute certainty which range the SZ patients fall within in terms of language comprehension, but the evidence presented here – from the pilot experiment and the previously cited literature (§5.7.1) – is suggestive of end5% type waveforms. SZ patients appear to elicit non-typical ERPs in response to certain aspects of language comprehension (i.e. the N400 and P600 ERPs as measures of meaning and syntax processing, respectively). If this is true, it would be consistent with the data outlined on language production in schizophrenia in §4.5, which was also suggestive of end5% behaviour. Therefore, on the PLTb profile of self, as it currently stands, SZ patients would be classified as possessing an abnormal form of self,

based on their language production and comprehension behaviour. This is consistent with other literature exploring the link between the self and schizophrenia (see Kircher & David, 2003; for overview), but the reasoning and motivation towards this conclusion is fundamentally different from these theories and is more robust to the type of criticisms outlined in §1.6.

One possible problem with this set of propositions is that SZ patients, themselves, are characteristically heterogeneous, in that their symptoms and behaviours vary significantly from patient to patient, but also within a patient (i.e. across time). Therefore, mass generalisations made on SZ patients will likely not be correct, for some patients will have a typical capacity for language. In lieu of this, it follows that the applicability of an abstract concept/noun criteria, established through PLTb theory, should be done on a case by case basis.

5.11 Chapter conclusion

This chapter has explored several aspects of language comprehension in humans, culminating in the presentation of two EEG experiments, which examine the neuronal processes underpinning this faculty. The first experiment was with neurotypical adults and examined the N400 and P600 ERPs as measures of semantic and syntactic processing, respectively. The results of this experiment indicated that in the comprehension of a written sentence, a violation in grammar correlates with a significant reduction in N400 amplitude on the critical word of the sentence/proposition, in comparison to a grammatical sentence. Furthermore, in the comprehension of grammatical sentences where the participant disagrees with its propositional meaning (i.e. finds the sentences to be false) there is an increase in N400 response relative to other sentence types (e.g. agree and ungrammatical sentences); a tentatively presented result, see §5.6.4 for reasons why. The other result of this experiment was related to methodology and led to the conclusion that caution is required when designing and analysing EEG experiments that examine the interplay between two (or more) ERPs, for the baseline selection for the later ERP can significantly affect the direction of the results.

The second experiment was with patients with a diagnosis of schizophrenia and neurotypical controls. This was a pilot experiment, which utilised the same task as the first experiment. Given the low sample size and other confounds (outlined in §5.8), it was established that the ERP results of this pilot experiment must be interpreted with caution. However, there was a clear overall difference in N400 response between groups, with the controls eliciting a significantly increased N400 in comparison to the SZ patients. Further analysis revealed that verbal IQ and working memory capacity explain some of this variance.

Other conclusions of this pilot related to more practical elements, such as improving the experimental design (see §5.7.4 and 5.9).

In theoretical terms – those more related to the thesis argument – these experiments were conducted to address an evidence gap in the PLTb theory of self. Based on these experiments, and leaning heavily on the previous literature, it was concluded that: [1] the N400 and P600 reliably index the processing of meaning and syntax, respectively, and that [2] patients with a diagnosis of schizophrenia elicit an abnormal N400 response in comparison the neurotypical adults. In view of [1], it was proposed that these components (i.e. their associated neuronal behaviour) be integrated into the PLTb theory of self, as representative of comprehension behaviour. Therefore, the PLTb theory of self, as it currently stands – which as discussed previously, is 'work in progress' – has language production (§4.5) and language comprehension behaviours serving as criteria for the typical human self, with the inclusion/exclusion, criterial threshold dictated by the statistical model of normal distribution. With this theory in place, and given the behaviours exhibited by patients with a diagnosis of schizophrenia, in both production and comprehension, it seems that they would not satisfy the criteria for typical selfhood. This view is consistent with the previous literature, but is motivated by a completely different framework (i.e. PLTb theory), which is less susceptible to the criticisms and linguistic biases outlined in Chapter 1 and 3, respectively.

6 THESIS SUMMARY AND CONCLUSIONS

This final chapter summarises the important ideas, arguments, analyses and conclusions established in this thesis. To an extent, the content provided here mirrors that presented in the Preface, but in this reflection, the summary points are linked to possible future avenues of research.

Chapter 1 reviewed sixteen theories of the self, from philosophy, psychology and cognitive neuroscience. Based on this review, it was determined that when these individual theories of self are considered as standalone, independent theories, they often prove to be internally consistent and persuasive, but when these theories are considered as a set of theories, significant contradictions arise, which undermines the current approach to the self and to other abstract noun/concepts. This chapter motivated a step-back from the debate of self, to consider more foundational and metatheoretical areas relating to nouns/concept acquisition and structure.

In future research, it would be beneficial to extend the review to other theories of self, to further demonstrate the above point of contradiction. It would also be useful to develop case studies of other abstract nouns/concepts (e.g. 'consciousness', 'justice', 'morality'), and complementary to this, it would be valuable to perform an etymological examination of the self and other key nouns/concepts, to understand their origins and the ways in which their meanings have evolved over time.

Chapter 2 examined how the meanings associated with concrete and abstract nouns arise in humans. It was argued that they arise in different ways and are subject to different cognitive processes. Concrete nouns are typically acquired ostensively, through perceptuo-linguistic triangulation (PLT), where there is a three-fold, cross-modal, perceptual interaction between two or more individuals and an object in the world. Abstract nouns, devoid of a clear third-point physical object in the world, are acquired through point-to-point (P2P) interactions between two individuals, with one giving the information and one receiving the information. However, this P2P information can be supplemented by a form of triangulation involving two or more individuals and a reference to a behaviour, this was termed: 'perceptuo-linguistic triangulation of behaviour' (PLTb).

With a basic understanding of how concrete and abstract noun meanings arise, the discussion moved on to understanding their underlying meaning structure. This started with a review and critique of the classical explanation of noun/concept structure. Several problems were highlighted, leading to the conclusion that this is an inadequate approach to explaining

concepts and solving their associated problems. The critique of the classical theory was followed by a description and analysis of prototype theory, as an alternative, statistical, approach to explaining word/noun/concept structure. It was argued that this theory better explains the structuring of meanings, because it accounts for, and predicts, the empirical data, as well as satisfying important theoretical requirements. After establishing prototype theory as a viable explanation of meaning structure, the nature of concrete and abstract nouns were described from within this perspective. It is concluded that concrete nouns have stable meanings, because the objects they name have a shared existence in the perceptual domain and are thus susceptible to PLT; PLT being the stabilising force. Whereas, abstract nouns are unstable, since they name things lacking a shared existence in the perceptual domain, and thus are not susceptible to PLT. In spite of this difference, both of these type of nouns can occupy similar grammatical-referential structures.

The next stage of this chapter explored Fodor's (1998) compositionality principle, as a potential barrier to prototype theory. It was argued that compositionality is a misguided principle, for it cannot account for referential (and propositional) meaning, and that it is only through grammar, that referential meaning is possible. In the final section of this chapter, this grammatical conception of referential meaning was linked to Vosse and Kempen's (2000) computational parsing theory of language (their 'u-model'). It was argued that this combined theory of grammar and parsing jointly explain important aspects of referential meaning; described in detail in Chapter 3.

An important research avenue that presents itself from the content of Chapter 2 relates to the topic of language acquisition. More specifically, further theoretical and empirical research is required on how nouns and other words are acquired in children. Furthermore, once these words are established, it is of interest to know how the brain – perhaps via a similarity-based mechanism, such as that described in prototype theory – facilitates classification decisions in the here-and-now? What are the neuronal mechanisms that make this possible?

Chapter 3 argued that certain forms of theorising – those that explicitly or implicitly assume aspects of the classical theory, described in Chapter 2 – lead to a specific type of ontological misclassification, termed 'noun phase (NP) reification', and to subtle, unintended changes in propositional meaning, termed 'clausal reification'. The former, NP reification, involves a specific type of a Rylean (1949/2009) category mistake in which an abstract noun is mistakenly taken to be of the concrete category. The latter, clausal reification, occurs when

an abstract noun, embedded in an NP, occupies the subject position of a proposition, which is making an essentialist claim about an abstract noun/concept.

NP and clausal reification are framed as linguistic artefacts/biases, as unintentional products of the way human language functions. It was hypothesised that they occur for reasons relating to the construction of referential meaning. Building on the work in Chapter 2, it was argued that referential meaning is a grammatical phenomenon, and that different grammatical relations correlate with different referential strengths, which can be captured in a grammatical, referential hierarchy. When this hierarchy is considered in conjunction with Vosse and Kempen's u-model, they jointly predict that a noun's lexical content (its meaning) is blindly integrated into already-established referential structure. Therefore, abstract nouns can enter referential configurations that are ill suited for their fuzzy level of meaning stability. This mismatch is the root cause of both NP and clausal reification. On this point, it was concluded that the existence of these biases explains the diversity of seemingly plausible but contradictory positions documented on the self in Chapter 1.

The final part of this chapter positioned these biases in the wider context of the cognitive bias literature. This literature holds that the human form of cognition is prone to make systematic errors. To date, language and grammar, as they are conceptualised in this thesis, have not been considered as an area vulnerable to cognitive biases. This discussion constituted an attempt to begin this conversation.

It was suggested that NP and clausal reification should be framed as linguistic biases and embedded in the cognitive bias literature. But the cognitive biases literature typically leans heavily on empirical evidence. Therefore, to develop the ideas of NP and clausal reification, it is important to try to think about to what extent these ideas make predictions that can be empirically investigated. It would also be beneficial to try to identify if there are any other grammatical, linguistic biases similar to the ones described in this thesis. An examination of languages other than English could feed into both of these proposed research avenues.

Chapter 4 advanced a positive method for theorising on abstract nouns/concepts, demonstrated with reference to the self. In this chapter (and in Chapter 3), it was argued that abstract nouns need to be stabilised for them to be both useful and synchronised (in their use) across a linguistic community. By default, PLT was considered an inappropriate stabilising force, because abstract nouns are devoid of a clear third-point physical object in the world that would facilitate triangulation. However, it was argued that PLTb may be an appropriate

alternative, for abstract nouns can be linked to behaviours, which have an observable, shared existence in the perceptual domain and thus are susceptible to triangulation. This appeal to behaviour was advanced as an antidote to classical, essentialist theorising, which, as argued in Chapter 3, leads to quasi-tautological propositions and theories.

It was established that one of the major difficulties in utilising this PLTb approach (to the self and other abstract concepts) was isolating the behaviours of interest without resorting to the essentialist approach. Using PLTb, in conjunction with basic scientific methods and the Gaussian probability distribution in statistics – which led to the mid95% and end5% classification ideas – it was possible to circumvent this trap.

With this theory in place, the argumentation moved on to a discussion of which behaviours to integrate into the PLTb profile of self. In this first instance, the first person pronoun 'I' was considered, but was found to be inadequate. In its place, the production of fully socialised and statistically typical utterances were proposed as a core constituent behaviour in the profile of self

Next, possible objections to utilising this PLTb method were discussed. The last criticism outlined, on artificial intelligence, highlighted the need to integrate behaviour associated with human comprehension into the PLTb profile of self, that is, in addition to the production of utterances, discussed in the previous paragraph. It was proposed that this would be achieved via the triangulation of neuronal behaviour associated with the N400 and P600, electroencephalography (EEG) event related potentials (ERPs). Evidence suggests that these ERPs index semantic and syntactic comprehension, respectively. This discussion primed the empirical investigations in Chapter 5.

Chapter 4 represents the beginnings of a new approach to theorising on the self and other abstract nouns/concepts. Significant work is still required to develop the details of this theory and more so in applying it to other nouns/concepts. The PLTb theory of self, as it currently stands, is linked to behaviours associated with utterance production and neuronal language comprehension, but further integration of other relevant behaviours is required before this notion could be factored in to important debates that lean on the notion of self. As discussed in Chapter 4, it is hoped that this will be a collaborative effort between researchers, rather than an individualistic enterprise, for the latter can encourage a competitive and uncooperative spirit.

When further developed, PLTb theory should be applied to other nouns/concepts, but with prioritisation given to nouns/concepts where a PLTb classification distinction (i.e. mid95% or end5%) would lead to different and beneficial actions and outcomes. For example, in developing a PLTb theory of self, it was observed the patients with a diagnosis of schizophrenia may not satisfy the PLTb criteria of selfhood, in that they appear to demonstrate statistically abnormal (end5%) utterance productions and neuronal language comprehension behaviour. It would be especially advantageous if once this end5% information is established, interventions could be developed and employed to help adjust these behaviours, that is, **if and only if** the individual finds the behaviour to be disruptive in their lives. The identification of end5% behaviour in utterance production and neuronal comprehension in schizophrenia could lead to the development of interventional therapies targeted at these areas. For example, the development of speech and language therapies for the production aspect and the employment of neurostimulation techniques for the neuronal, comprehension aspect. However, significantly-more research would be required in both of these suggested areas before these interventions were possible; especially in the latter, for neurostimulation techniques incorrectly employed could present serious risks to health.

One noun/concept where the application of PLTb theory may be especially fruitful is in the area of disease, for much work, but little progress has been made on this concept. The development of a PLTb approach to disease – which would sync neatly with Boorse's (1977) biostatical theory of disease – could potentially help in clinical decision making and the development of diagnostic devices. The connection between PLTb classifications derived from Gaussian probability distribution and the decision making or actionable outcomes (mentioned above) could potentially be mapped via fuzzy logic (Zadeh, 1965). Fuzzy logic is a mechanism that imitates human decision making, where there are defined possibilities that fall between the binary values of 'yes' and 'no'. In short, there is strong connection between fuzzy logic, prototype theory and probability distributions, which could aid in the development of PLTb theory.

There is a strong Wittgensteinian influence operating (mostly) behind the scenes of this thesis. Future philosophical research should be more explicit in exploring the relation between Wittgenstein's philosophy of language, and the things said on language in this thesis, which, as alluded to in §3.8, appear to straddle Wittgenstein's early work on picture theory and his later work on language games. Further work in this area will help place the ideas of this thesis within the wider context of historical and modern philosophical research.

Chapter 5 presented two EEG experiments that explored the neuronal processes underpinning the language comprehension mechanism, with a view to integrating this information – in conjunction with the previous literature – into the positive PLTb theory of self. This chapter began with an overview of the EEG technique, followed by a brief literature review of the N400 and P600 ERPs, which are associated with the processing of meaning and syntax, respectively. These ERPs were the focus of the two EEG experiments presented in this thesis.

The first experiment was with neurotypical adults, the second with patients with a diagnosis of schizophrenia and neurotypical adult controls. Patients with a diagnosis of schizophrenia were selected, because schizophrenia is often linked to a loss of self or attenuated selfhood. Therefore, it was of interest to explore to what extent that this is the case in view of this new PLTb approach to the self.

Overall, it was concluded that aspects of neurophysiological evidence (including EEG ERPs), where the components can be observed and measured statistically, can be integrated into a PLTb profile. Here, it was tentatively recommended that the N400 and P600 ERPs be integrated into the PLTb profile of self as triangulable measures of language comprehension.

The final section of this chapter briefly discussed the consequences of this integration in the context of schizophrenia. As mentioned above, it was concluded that, based on the PLTb theory of self, as it currently stands, patients with a diagnosis of schizophrenia would not likely satisfy the PLTb criteria of selfhood, in that they appear to demonstrate statistically abnormal (end5%) neuronal language comprehension behaviours.

Future directions in this experimental area are extensively considered in §5.9, so will not be repeated here. Besides, to say that it is hoped that after this research, this marriage between philosophical theorising and empirical investigation can continue to grow, for work performed on the fringes of these two disciplines, where they intersect and overlap, generates research questions and solutions that may not be visible to the mind of the ultra-specialised, siloed thinker.

7 REFERENCES

Abney, S. P. (1987). *The English noun phrase in its sentential aspect.* Massachusetts Institute of Technology.

Aerts, D. (2014). Quantum and Concept Combination, Entangled Measurements, and Prototype Theory. *Topics in Cognitive Science, 6*(1), 129-137.

Aerts, D., Broekaert, J., Gabora, L., & Sozzo, S. (2016). Generalizing Prototype Theory: A Formal Quantum Framework. *Front Psychol, 7*(418).

Ainsworth-Darnell, K., Shulman, H. G., & Boland, J. E. (1998). Dissociating Brain Responses to Syntactic and Semantic Anomalies: Evidence from Event-Related Potentials. *Journal of Memory and Language, 38*(1), 112-130.

Alschuler, D. M., Tenke, C. E., Bruder, G. E., & Kayser, J. (2014). Identifying electrode bridging from electrical distance distributions: a survey of publicly-available EEG data using a new method. *Clinical neurophysiology: official journal of the International Federation of Clinical Neurophysiology, 125*(3), 484-490.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

Aristotle. (384-322BCE/1933). Metaphysics (H. Tredennick, Trans.). London: Heinemann.

Aristotle. (384-322BCE/2016). *Aristotle: De Anima* (C. Shields, Trans.). Oxford: Clarendon Press.

Augustine. (354-430/2003). *The City of God* (H. Bettenson, Trans.). London: Penguin Books.

Austin, J. L. (1962). *How to Do Things with Words* (2nd ed.). Cambridge: Harvard University Press.

Barresi, J., & Martin, R. (2013). History as prologue: western theories of the self. In S. Gallagher (Ed.), *The Oxford handbook of the self* (pp. 57-80). Oxford: Oxford University Press.

Barton, S. B., & Sanford, A. J. (1993). A case study of anomaly detection: shallow semantic processing and cohesion establishment. *Memory & Cognition, 21*(4), 477-487.

Bastiaansen, M., & Hagoort, P. (2015). Frequency-based segregation of syntactic and semantic unification during online sentence level language comprehension. *J Cogn Neurosci, 27*(11), 2095-2107.

Benjamin, L. T., Jr. (2000). The psychology laboratory at the turn of the 20th century. *American Psychologist, 55*(3), 318-321.

Birn, R. M., Bandettini, P. A., Cox, R. W., & Shaker, R. (1999). Event-related fMRI of tasks involving brief motion. *Hum Brain Mapp, 7*(2), 106-114.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, Massachusetts: MIT Press.

Bloom, P. (2001). Precis of How children learn the meanings of words. *The Behavioral and Brain Sciences, 24*(6).

Bongard, J., Zykov, V., & Lipson, H. (2006). Resilient machines through continuous self-modeling. *Science, 314*(5802), 1118-1121.

Boorse, C. (1977). Health as a theoretical concept. *Philosophy of Science, 44*(4), 542-573.

Botvinick, M., & Cohen, J. (1998). Rubber hands' feel'touch that eyes see. *Nature, 391*(6669), 756.

Bright, P., Hale, E., Gooch, V. J., Myhill, T., & van der Linde, I. (2018). The National Adult Reading Test: restandardisation against the Wechsler Adult Intelligence Scale-Fourth edition. *Neuropsychol Rehabil, 28*(6), 1019-1027.

Brook, A. (2016). Kant's View of the Mind and Consciousness of Self. In E. N. Zalta (Ed.), *Standford Encyclopedia of Philosophy*.

Brown, G. D., Yamada, S., & Sejnowski, T. J. (2001). Independent component analysis at the neural cocktail party. *Trends Neurosci, 24*(1), 54-63.

Butler, J. (1975). Of Personal Identity. In J. Perry (Ed.), *Personal Identity* (pp. 99-105). California: Univeristy of California Press.

Campbell, J. (1999). Schizophrenia, The Space of Reasons, and Thinking as a Motor Process. *The Monist, 82*(4), 609-625.

Caplan, D., Michaud, J., Hufford, R., & Makris, N. (2016). Deficit-lesion correlations in syntactic comprehension in aphasia. *Brain and Language, 152*, 14-27.

Cappelen, H., & Dever, J. (2013). *The Inessential Indexical: On the Philosophical Insignificance of Perspective and the First Person*: Oxford University Press.

Carlson, G. N. (1977). A unified analysis of the English bare plural. *Linguistics and Philosophy, 1*(3), 413-457.

Carlson, G. N. (1980). Reference to kinds in English. *PhD thesis, University of California*.

Chakrabarti, K. (1975). The Nyāya-Vaiśeṣika theory of universals. *Journal of Indian Philosophy, 3*(3), 363-382.

Chomsky, N. (1957/2002). *Syntactic Structures* (2nd ed.). Berlin: Walter de Gruyter.

Chomsky, N. (1995/2014). *The Minimalist Program*. Cambridge, Massachusetts: MIT Press.

Churchland, P. M. (1981). Eliminative Materialism and the Propositional Attitudes. *The Journal of Philosophy, 78*(2), 67-90.

Çokal, D., Sevilla, G., Jones, W. S., Zimmerer, V., Deamer, F., Douglas, M., . . . Hinzen, W. (2018). The language profile of formal thought disorder. *NPJ Schizophr, 4*(1), 18.

Coles, M. G. H., Gratton, G., Kramer, A., & Miller, G. A. (1986). Principles of signal acquisition and analysis. In M. G. H. Coles, E. Donchin, & S. W. Porges (Eds.), *Psychophysiology: Systems, processes, and applications*. New York: Guilford Press.

Coulson, S., King, J. W., & Kutas, M. (1998). Expect the Unexpected: Event-related Brain Response to Morphosyntactic Violations. *Language and cognitive processes, 13*(1), 21-58.

Craik, F. I. M., Moroz, T. M., Moscovitch, M., Stuss, D. T., Winocur, G., Tulving, E., & Kapur, S. (1999). In Search of the Self: A Positron Emission Tomography Study. *Psychological Science, 10*(1), 26-34.

Croft, R. J., & Barry, R. J. (2000). Removal of ocular artifact from the EEG: a review. *Neurophysiologie Clinique/Clinical Neurophysiology, 30*(1), 5-19.

Cutler, A., & Clifton, C. E. (1999). Comprehending spoken language: A blueprint of the listener. In C. M. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 123-166). Oxford, UK: Oxford University Press.

Dąbrowska, E. (2015). What exactly is Universal Grammar, and has anyone seen it? *Front Psychol, 6*(852).

Davidson, D. (1982). Rational Animals. *Dialectica, 36*(4), 317-327.

Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning: neural and behavioural evidence. *Philosophical Transactions of the Royal Society B, 364*(1536), 3773-3800.

Dennett, D. C. (1992). The Self as a Center of Narrative Gravity. In F. S. Kessel, P. M. Cole, & D. L. Johnson (Eds.), *Self and Consciousness: Multiple Perspectives*. London: Lawrence Erlbaum Associates

Descartes, R. (1641/2008). *Meditations on first philosophy: With selections from the objections and replies* (M. Moriarty, Trans.). Oxford: Oxford University Press.

Devuyst, S., Dutoit, T., Stenuit, P., Kerkhofs, M., & Stanus, E. (2008). Removal of ECG artifacts from EEG using a modified independent component analysis approach. *Conf Proc IEEE Eng Med Biol Soc, 2008*, 5204-5207.

Dronkers, N. F., Plaisant, O., Iba-Zizen, M. T., & Cabanis, E. A. (2007). Paul Broca's historic cases: high resolution MR imaging of the brains of Leborgne and Lelong. *Brain, 130*(Pt 5), 1432-1441.

Dudschig, C., Maienborn, C., & Kaup, B. (2016). Is there a difference between stripy journeys and stripy ladybirds? The N400 response to semantic and world-knowledge violations during sentence processing. *Brain and Cognition, 103*, 38-49.

Duncan, C. C., Barry, R. J., Connolly, J. F., Fischer, C., Michie, P. T., Näätänen, R., . . . Van Petten, C. (2009). Event-related potentials in clinical research: Guidelines for eliciting, recording, and quantifying mismatch negativity, P300, and N400. *Clinical Neurophysiology, 120*(11), 1883-1908.

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior, 20*(5), 540-551.

Esslen, M., Metzler, S., Pascual-Marqui, R., & Jancke, L. (2008). Pre-reflective and reflective self-reference: a spatiotemporal EEG analysis. *Neuroimage, 42*(1), 437-449.

Evans, G. (1981). Understanding demonstratives. In H. Parret (Ed.), *Meaning and Understanding* (pp. 280-304): Clarendon Press.

Fara, D. G. (2015). Names are Predicates. *Philosophical Review, 124*(1), 59-117.

Farrer, C., Franck, N., Georgieff, N., Frith, C. D., Decety, J., & Jeannerod, M. (2003). Modulating the experience of agency: a positron emission tomography study. *Neuroimage, 18*(2), 324-333.

Farrer, C., & Frith, C. D. (2002). Experiencing Oneself vs Another Person as Being the Cause of an Action: The Neural Correlates of the Experience of Agency. *Neuroimage, 15*(3), 596-603.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191.

Federmeier, K. D., & Kutas, M. (2001). Meaning and modality: influences of context, semantic memory organization, and perceptual predictability on picture processing. *J Exp Psychol Learn Mem Cogn, 27*(1), 202-224.

Fiez, J. A. (2001). Neuroimaging studies of speech an overview of techniques and methodological approaches. *J Commun Disord, 34*(6), 445-454.

Fineberg, S. K., Deutsch-Link, S., Ichinose, M., McGuinness, T., Bessette, A. J., Chung, C. K., & Corlett, P. R. (2015). Word use in first-person accounts of schizophrenia. *The British Journal of Psychiatry, 206*(1), 32-38.

Fodor, J. A. (1998). *Concepts Where Cognitive Science Went Wrong*. Oxford: Clarendon Press.

Frege, G. (1892). Über Sinn und Bedeutung. *Zeitschrift für Philosophie Und Philosophische Kritik, 100*(1), 25--50.

Frege, G. (1918/1977). Thoughts. In P. T. Geach (Ed.), *Logical Investigations*. Oxford Basil, Blackwell. .

Friederici, A. D., Steinhauer, K., & Frisch, S. (1999). Lexical integration: sequential effects of syntactic and semantic information. *Mem Cognit, 27*(3), 438-453.

Friston, K. J., & Price, C. J. (2011). Modules and brain mapping. *Cognitive Neuropsychology, 28*(3-4), 241-250.

Frith, C. D. (1992). *The Cognitive Neuropsychology of Schizophrenia*. Hove: Lawrence Erlbaum Associates.

Frith, C. D. (2002). Attention to action and awareness of other minds. *Conscious Cogn, 11*(4), 481-487.

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends Cogn Sci, 4*(1), 14-21.

Gallagher, S. (2013). A pattern theory of self. *Front Hum Neurosci, 7*, 443.

Gergen, K. J. (2013). The Social Construction of Self. In S. Gallagher (Ed.), *The Oxford Handbook of The Self* (Papaerback edition. ed., pp. 633-653). Oxford: Oxford University Press.

Gettier, E. L. (1963). Is justified true belief knowledge? *analysis, 23*(6), 121-123.

Gillihan, S. J., & Farah, M. J. (2005). Is Self Special? A Critical Review of Evidence From Experimental Psychology and Cognitive Neuroscience. *Psychological Bulletin, 131*(1), 76-97.

Gleitman, L. R., Connolly, A. C., & Armstrong, S. L. (2012). Can prototype representations support compositions? . In M. Werning, W. Hinzen, & E. Machery (Eds.), *The Oxford handbook of compositionality* (pp. 418-436). Oxford Oxford University Press.

Gracco, V. L., Tremblay, P., & Pike, B. (2005). Imaging speech production using fMRI. *Neuroimage, 26*(1), 294-301.

Graham, G. (2017). "Behaviorism".

Groppe, D. M., Makeig, S., & Kutas, M. (2009). Identifying reliable independent components via split-half comparisons. *Neuroimage, 45*(4), 1199-1211.

Gunter, T. C., Stowe, L. A., & Mulder, G. (1997). When syntax meets semantics. *Psychophysiology, 34*(6), 660-676.

Hagoort, P. (2003a). How the brain solves the binding problem for language: a neurocomputational model of syntactic processing. *Neuroimage, 20, Supplement 1*, S18-S29.

Hagoort, P. (2003b). Interplay between syntax and semantics during sentence comprehension: ERP effects of combining syntactic and semantic violations. *J Cogn Neurosci, 15*(6), 883-899.

Hagoort, P. (2017). The core and beyond in the language-ready brain. *Neuroscience and Biobehavioural Reviews*.

Hagoort, P., & Brown, C. M. (1994). Brain responses to lexical ambiguity resolution and parsing *Perspectives on sentence processing* (pp. 45-80). Hillsdale, NJ.: Lawrence Erlbaum Associates, Inc.

Hagoort, P., Brown, C. M., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and cognitive processes, 8*(4), 439-483.

Hagoort, P., Brown, C. M., & Osterhout, L. (1999). The neurocognition of syntactic processing,. In C. M. Brown & P. Hagoort (Eds.), *Neurocognition of language* (pp. 273-317). Oxford: Oxford University Press.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science, 304*(5669), 438-441.

Hagoort, P., & van Berkum, J. (2007). Beyond the sentence given. *Philos Trans R Soc Lond B Biol Sci, 362*(1481), 801-811.

Hagoort, P., Wassenaar, M., & Brown, C. (2003). Real-time semantic compensation in patients with agrammatic comprehension: Electrophysiological evidence for multiple-route plasticity. *Proceedings of the National Academy of Sciences, 100*(7).

Hald, L. A., Bastiaansen, M. C., & Hagoort, P. (2006). EEG theta and gamma responses to semantic violations in online sentence processing. *Brain Lang, 96*(1), 90-105.

Hampton, J. A. (1995). Testing the Prototype Theory of Concepts. *Journal of Memory and Language, 34*(5), 686-708.

Hampton, J. A. (2011). Concepts and Natural Language. In R. Belohlavek & G. J. Klir (Eds.), *Concepts and Fuzzy Logic*: MIT Press.

Hanson, A. J. (2017). *Event-related EEG Analysis: Simple solutions or complex computations.* (PhD), Newcastle University.

Harrison, B. J., & Pantelis, C. (2010). Cognitive Subtraction. In I. P. Stolerman (Ed.), *Encyclopedia of Psychopharmacology*. Berlin, Heidelberg: Springer.

Hawking, S. (2001). *The Universe in a Nutshell*. London: Bantum Press.

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition, 92*(1-2), 67-99.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8*(5), 393-402.

Higgins, F. R. (1979). *The pseudo-cleft construction in English*. New York: Garland.

Hinzen, W. (2014). What is Un-Cartesian Linguistics? *Biolinguistics, 8*, 226-257.

Hinzen, W. (2016). Linguistic Evidence against Predicativism. *Philosophy Compass, 11*(10), 591-608.

Hinzen, W., Rossello, J., & McKenna, P. (2016). Can delusions be understood linguistically? *Cogn Neuropsychiatry, 21*(4), 281-299.

Hinzen, W., & Sheehan, M. (2013). *The Philosophy of Universal Grammar*. Oxford: Oxford University Press.

Holmberg, A. (2010). How to refer to yourself when talking to yourself. *Newcastle Working Papers in Linguistics 16*.

Hume, D. (1739/1888). *A Treatise of Human Nature* (L. A. Selby-bigge Ed.). Oxford: Clarendon Press.

Humstone, H. J. (1919). Memory Span Tests. *Psychol. Clin., 12*, 196-200.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature, 532*(7600), 453.

Indefrey, P. (2004). Hirnaktivierungen bei syntaktischer sprachverarbeitung: Eine meta-analyse. In H. Müller & G. Rickheit (Eds.), *Neurokognition der Sprache* (pp. 31-50): Stauffenburg.

Indefrey, P., & Cutler, A. (2004). Prelexical and lexical processing in listening. In M. Gazzaniga (Ed.), *The cognitive neurosciences III.* (pp. 759-774). Cambridge: MIT Press.

Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological Science, 16*(5), 367-371.

Izard, C. E. (1972). *Patterns of Emotions: A New Analysis of Anxiety and Depression*. New York: Academic Press.

Jackson, F., Foti, D., Kotov, R., Perlman, G., Mathalon, D. H., & Proudfit, G. H. (2014). An Incongruent Reality: The N400 in Relation to Psychosis and Recovery. *Schizophrenia Research, 160*(0), 208-215.

James, W. (1890/2007). *The Principles of Psychology, Vol. 1*. New York: Cosimo.

Johnson, S. C., Baxter, L. C., Wilder, L. S., Pipe, J. G., Heiserman, J. E., & Prigatano, G. P. (2002). Neural correlates of self-reflection. *Brain, 125*(Pt 8), 1808-1814.

Jones, S. S., Smith, L. B., & Landau, B. (1991). Object properties and knowledge in early lexical learning. *Child Dev, 62*(3), 499-516.

Kahneman, D. (2011). *Thinking, Fast and Slow*. London: Penguin Books.

Kampe, K. K., Frith, C. D., & Frith, U. (2003). "Hey John": signals conveying communicative intention toward the self activate brain regions associated with "mentalizing," regardless of modality. *The Journal of Neuroscience, 23*(12), 5258-5263.

Kanner, L. (1943). Autistic disturbances of affective contact. *Nervous child, 2*(3), 217-250.

Kant, I. (1781/2007). *Critique of Pure Reason* (M. Weigelt, Trans. M. Weigelt Ed.). London: Penguin Books.

Kaplan, D. (1989). Demonstratives. In J. Almog, J. Perry, & H. Wettstein (Eds.), *Themes From Kaplan* (pp. 481-563). Oxford: Oxford University Press.

Kappenman, E. S., & Luck, S. J. (2010). The Effects of Electrode Impedance on Data Quality and Statistical Significance in ERP Recordings. *Psychophysiology, 47*(5), 888-904.

Kass, R. E., Caffo, B. S., Davidian, M., Meng, X. L., Yu, B., & Reid, N. (2016). Ten Simple Rules for Effective Statistical Practice. *PLoS Comput Biol, 12*(6).

Kay, P., & McDaniel, C. K. (1978). The linguistic significance of the meanings of basic color terms. *Language*, 610-646.

Kelley, W. M., Macrae, C. N., Wyland, C. L., Caglar, S., Inati, S., & Heatherton, T. F. (2002). Finding the Self? An Event-Related fMRI Study. *J Cogn Neurosci, 14*(5), 785-794.

Kempen, G., & Vosse, T. (1989). Incremental Syntactic Tree Formation in Human Sentence Processing: a Cognitive Architecture Based on Activation Decay and Simulated Annealing. *Connection Science, 1*(3), 273-290. doi:10.1080/09540098908915642

Kiang, M., Kutas, M., Light, G. A., & Braff, D. L. (2007). Electrophysiological insights into conceptual disorganization in schizophrenia. *Schizophrenia Research, 92*(0), 225-236.

Kiran, C., & Chaudhury, S. (2009). Understanding delusions. *Industrial Psychiatry Journal, 18*(1), 3-18.

Kircher, T., & David, A. (Eds.). (2003). *THE SELF in Neuroscience and Psychiatry*. Cambridge: Cambridge University Press.

Kircher, T., Senior, C., Phillips, M. L., Rabe-Hesketh, S., Benson, P. J., Bullmore, E. T., . . . David, A. S. (2001). Recognizing one's own face. *Cognition, 78*(1), B1-B15.

Kirkbride, J. B., Errazuriz, A., Croudace, T. J., Morgan, C., Jackson, D., Boydell, J., . . . Jones, P. B. (2012). Incidence of Schizophrenia and Other Psychoses in England, 1950–2009: A Systematic Review and Meta-Analyses. *PLoS ONE, 7*(3).

Kirschstein, T., & Köhling, R. (2009). What is the source of the EEG? *Clinical EEG and neuroscience, 40*(3), 146-149.

Klem, G. H., Luders, H. O., Jasper, H. H., & Elger, C. (1999). The ten-twenty electrode system of the International Federation. The International Federation of Clinical Neurophysiology. *Electroencephalogr Clin Neurophysiol Suppl, 52*, 3-6.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (Vol. 2). Chicago: International Encyclopedia of Unified Science.

Kumar, N., & Debruille, J. B. (2004). Semantics and N400: insights for schizophrenia. *Journal of psychiatry & neuroscience, 29*(2), 89.

Kutas, M., & Federmeier, K. D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP) *Annu. Rev. Psychol.* (Vol. 62, pp. 621-647).

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science, 207*(4427), 203-205.

Kutas, M., Neville, H. J., & Holcomb, P. J. (1987). A preliminary comparison of the N400 response to semantic anomalies during reading, listening and signing. *Electroencephalogr Clin Neurophysiol Suppl, 39*, 325-330.

Labov, W. (1973). The boundaries of words and their meanings. In C. J. N. Bailey & R. W. Shuy (Eds.), *New Ways of Analysing Variation in English* (pp. 340-373). Washington: Georgetown University Press.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Front Psychol, 4*, 863.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago: University of Chicago Press.

Lam, N. H. L., Schoffelen, J.-M., Uddén, J., Hultén, A., & Hagoort, P. (2016). Neural activity during sentence processing as reflected in theta, alpha, beta, and gamma oscillations. *Neuroimage*. doi:10.1016/j.neuroimage.2016.03.007

Landau, B., Smith, L., & Jones, S. S. (1998). Object Shape, Object Function, and Object Name. *Journal of Memory and Language, 38*(1), 1-27.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The Importance of Shape in Early Lexical Learning. *Cognitive development, 3*(3), 299-321.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2-3), 259-284.

Lattner, S., & Friederici, A. D. (2003). Talker's voice and gender stereotype in human auditory sentence processing – evidence from event-related brain potentials. *Neuroscience Letters, 339*(3), 191-194.

Leary, D. (1990). William James on the Self and Personality: Clearing the Ground for Subsequent Theorists, Researchers, and Practitioners. In M. G. Johnson & T. B. Henley (Eds.), *Reflections on The Principles of Psychology: William James after a Century* (pp. 101-137). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Legrand, D., & Ruby, P. (2009). What Is Self-Specific? Theoretical Investigation and Critical Review of Neuroimaging Results. *Psychological Review, 116*(1), 252-282.

Locke, J. (1689/1975). *An Essay Concerning Human Understanding* (P. H. Nidditch Ed.). Oxford: Oxford University Press.

Longobardi, G. (1994). Reference and Proper Names: A Theory of N-Movement in Syntax and Logical Form. *Linguistic Inquiry, 25*(4), 609-665.

Longobardi, G. (2005). Toward a Unified Grammar of Reference *Zeitschrift für Sprachwissenschaft* (Vol. 24, pp. 5).

Luck, S. J. (2005). *An introduction to the event-related potential technique*. Cambridge, Mass.: MIT Press.

Maatz, A. (2014). Use of the first-person pronoun in schizophrenia. *Br J Psychiatry, 205*(5), 409.

Macrae, C. N., Moran, J. M., Heatherton, T. F., Banfield, J. F., & Kelley, W. M. (2004). Medial Prefrontal Activity Predicts Memory for Self. *Cerebral Cortex, 14*(6), 647-654.

Magidor, O. (2013). *Category Mistakes*: Oxford University Press.

Makeig, S., Jung, T.-P., Bell, A. J., Ghahremani, D., & Sejnowski, T. J. (1997). Blind separation of auditory event-related brain responses into independent components. *Proceedings of the National Academy of Sciences of the United States of America, 94*(20), 10979-10984.

Maravita, A., & Iriki, A. (2004). Tools for the body (schema). *Trends Cogn Sci, 8*(2), 79-86.

Marshall, J. C., & Fink, G. R. (2001). Spatial cognition: where we were and where we are. *Neuroimage, 14*(Pt 2), S2-S7.

Martin, R., & Barresi, J. (2006). *The Rise and Fall of Soul and Self: An Intellectual History of Personal Identity*: Columbia University Press.

Martin, T., & Hinzen, W. (2014). The Grammar of the Essential Indexical. *Lingua: International Review of General Linguistics, 148*, 95-117.

McGinn, C. (2016). *Philosphy of Language: The classics explained*. Cambridge, Massachusetts: MIT Press.

McNally, L., & Van, G. V. (1998). *Redefining the weak/strong distinction.* Paper presented at the Paris Syntax and Semantics Colloquium.

McRobbie, D. W., Moore, E. A., Graves, M. J., & Prince, M. R. (2006). *MRI From Picture to Proton*. Cambridge: Cambridge University Press.

Melnick, A. (2009). *Kant's Theory of Self*. London: Routledge.

Metzinger, T. (2003). *Being No One: The Self-Model Theory of Subjectivity*. Cambridge, Massachusetts: MIT Press.

Metzinger, T. (2005). Out-of-body experiences as the origin of the concept of a 'soul '. *Mind and Matter, 3*(1), 57-84.

Metzinger, T. (2008). Empirical perspectives from the self-model theory of subjectivity: a brief summary with examples. *Prog Brain Res, 168*, 215-245.

Metzler, S., Theodoridou, A., Aleksandrowicz, A., Müller, M., Obermann, C., Kawohl, W., & Heekeren, K. (2014). Evaluation of trait adjectives and ego pathology in schizophrenia: An N400 study. *Psychiatry research, 215*(3), 533-539.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *J Exp Psychol, 90*(2), 227-234.

Mikkelsen, L. (2011). Copular Clauses. In C. Maienborn, K. von Heusinger, & P. Portner (Eds.), *Semantics: An International Handbook of Natural Language Meaning* (pp. 1805-1829): De Gruyter Mouton.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81.

Millett, D. (2001). Hans Berger: From Psychic Energy to the EEG. *Perspectives in Biology and Medicine, 44*(4), 522-542.

Milsark, G. (1977). Toward an explanation of certain peculiarities of the existential construction in English. *Linguistic Analysis, 3*, 1-29.

Mohammad, O. M., & DeLisi, L. E. (2013). N400 in schizophrenia patients. *Curr Opin Psychiatry, 26*(2), 196-207.

Naigles, L. R., Cheng, M., Xu, R. N., Tek, S., Khetrapal, N., Fein, D., & Demuth, K. (2016). "You're telling me!" The prevalence and predictors of pronoun reversals in children with autism spectrum disorders and typical development. *Research in Autism Spectrum Disorders, 27*, 11-20.

Neisser, U. (1988). Five kinds of self-knowledge. *Philosophical Psychology, 1*(1), 35-59.

Nelson, H. E. (1982). *The National Adult Reading Test*. Windsor, UK.: NFER-Nelson.

Nieuwland, M. S., & Van Berkum, J. J. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *J Cogn Neurosci, 18*(7), 1098-1111.

Northoff, G. (2005). Emotional-cognitive integration, the self, and cortical midline structures. *Behavioral and Brain Sciences, 28*(2), 211-212.

Northoff, G. (2011). Self and brain: what is self-related processing? *Trends Cogn Sci, 15*(5), 186-187.

Northoff, G. (2013). Brain and self - a neurophilosophical account. *Child and Adolescent Psychiatry and Mental Health, 7*(28), 1-12.

Northoff, G., & Bermpohl, F. (2004). Cortical midline structures and the self. *Trends Cogn Sci, 8*(3), 102-107.

Northoff, G., Heinzel, A., de Greck, M., Bermpohl, F., Dobrowolny, H., & Panksepp, J. (2006). Self-referential processing in our brain--a meta-analysis of imaging studies on the self. *Neuroimage, 31*(1), 440-457.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: learning, memory, and cognition, 14*(4), 700.

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. *Essays in honor of William K. Estes, 1*, 149-167.

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychon Bull Rev, 7*(3), 375-402.

Nunez, P. L., & Srinivasan, R. (2006). *Electric fields of the brain: the neurophysics of EEG*. Oxford: Oxford University Press.

Ochsner, K. N., Beer, J. S., Robertson, E. R., Cooper, J. C., Gabrieli, J. D. E., Kihsltrom, J. F., & Esposito, M. (2005). The neural correlates of direct and reflected self-knowledge. *Neuroimage, 28*(4), 797-814.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia, 9*(1), 97-113.

Olson, E. T. (1998). There is no problem of the self. *Journal of Consciousness Studies, 5*(5-6), 645-657.

Onitsuka, T., Oribe, N., Nakamura, I., & Kanba, S. (2013). Review of neurophysiological findings in patients with schizophrenia. *Psychiatry Clin Neurosci, 67*(7), 461-470.

Ortu, D. (2012). *Exploring the Nature of Neural Correlates of Language, Attention and Memory: Reliability and Validity Studies of Event Related Potentials.* (PhD), University of Sterling.

Osherson, D. N., & Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition, 9*(1), 35-58.

Osterhout, L. (1997). On the brain response to syntactic anomalies: manipulations of word position and word class reveal individual differences. *Brain Lang, 59*(3), 494-522.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language, 31*(6), 785-806.

Osterhout, L., Holcomb, P. J., & Swinney, D. A. (1994). Brain potentials elicited by garden-path sentences: evidence of the application of verb information during parsing. *J Exp Psychol Learn Mem Cogn, 20*(4), 786-803.

Osterhout, L., & Nicol, J. (1999). On the distinctiveness, independence, and time course of the brain responses to syntactic and semantic anomalies. *Language and cognitive processes, 14*(3), 283-317.

Overall, J. E., & Gorham, D. R. (1962). The Brief Psychiatric Rating Scale. *Psychological Reports, 10*, 799-812.

Perrin, F., Maquet, P., Peigneux, P., Ruby, P., Degueldre, C., Balteau, E., . . . Laureys, S. (2005). Neural mechanisms involved in the detection of our first name: a combined ERPs and PET study. *Neuropsychologia, 43*(1), 12-19.

Perry, J. (1979). The Problem of the Essential Indexical. *Nous, 13*(1), 3-21.

Phan, K. L., Wager, T. D., Taylor, S. F., & Liberzon, I. (2004). Functional Neuroimaging Studies of Human Emotions. *CNS Spectr, 9*(4), 258-266.

Plato. (429-347BCE/1997). Plato Complete Works. In J. M. Cooper & D. S. Hutchinson (Eds.). Indianapolis: Hacket Publishing Company.

Popper, K. R. (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge*. New York: Basic Books.

Popper, K. R. (1980). *The Logic of Scientific Discovery*. Tiptree: The Anchor Press.

Pryor, J. (1999). Immunity to error through misidentification. *Philosophical Topics, 26*(1/2), 271-304.

Raichle, M. E. (1987/2011). Circulatory and metabolic correlates of brain function in normal humans. *Comprehensive Physiology*.

Ramachandran, V. S., & Rogers-Ramachandran, D. (1996). Synaesthesia in phantom limbs induced with mirrors. *Proc Biol Sci, 263*(1369), 377-386.

Ramsey, F. P. (1927). Facts and Propositions. *Proceedings of the Aristotelian Society, 7*(1), 153-170.

Reid, T. (1850). *Essays on the intellectual powers of man* (J. Walker Ed.). Cambridge: J. Bartlett.

Rochat, P. (2013). What is it like to be a Newborn? In S. Gallagher (Ed.), *The Oxford handbook of the self* (Papaerback edition. ed., pp. 57-80). Oxford: Oxford University Press.

Rosch, E. (1973). Natural categories. *Cognitive Psychology, 4*(3), 328-350.

Rosch, E. (1975a). Cognitive reference points. *Cognitive Psychology, 7*(4), 532-547.

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

Rosch, E. (1977). Classification of real-world objects: Origins and representations in cognition. *Thinking: Readings in cognitive science*, 212-222.

Rosch, E. (1999). Reclaiming concepts. *Journal of Consciousness Studies, 6*(11-12), 61-77.

Rosch, E., & Mervis, C. B. (1975). Family Resemblance: Studies in the Internal Structure of Categories. *Cognitive Psychology, 7*, 573-605.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*(3), 382-439.

Ross, J. R. (1970). On Declarative Sentences. In R. A. Jacobs & P. S. Rosenbaum (Eds.), *Readings in English transformational grammar*: Ginn.

Ryle, G. (1949/2009). *The Concept of Mind*. Oxford: Routledge.

Salisbury, D. F. (2008). Semantic activation and verbal working memory maintenance in schizophrenic thought disorder: insights from electrophysiology and lexical ambiguity. *Clin EEG Neurosci, 39*(2), 103-107.

Salisbury, D. F. (2010). N400 to lexical ambiguity and semantic incongruity in schizophrenia. *International Journal of Psychophysiology, 75*(2), 127-132.

Sanford, A. J. (2002). Context, attention and depth of processing during interpretation. *Mind & Language, 17*(1-2), 188-206.

Sass, L. A., & Parnas, J. (2003). Schizophrenia, Consciousness, and the Self. *Schizophrenia Bulletin, 29*(3), 427-444.

Sass, L. A., Pienkos, E., Nelson, B., & Medford, N. (2013). Anomalous self-experience in depersonalization and schizophrenia: A comparative investigation. *Conscious Cogn, 22*(2), 430-441.

Scott-Phillips, T. (2014). *Speaking Our Minds: Why human communication is different, and how language evolved to make it special*: Palgrave MacMillan.

Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.

Searle, J. R. (1980). Minds, brains and programs. *Behavioral and Brain Sciences, 3*(3), 417-457.

Semlitsch, H. V., Anderer, P., Schuster, P., & Presslich, O. (1986). A solution for reliable and valid reduction of ocular artifacts, applied to the P300 ERP. *Psychophysiology, 23*(6), 695-703.

Sheehan, M., & Hinzen, W. (2011). Moving Towards the Edge. *Linguistic Analysis, 37*(3-4), 1-54.

Shoemaker, S. (1984). *Identity, Cause and Mind*: Cambridge University Press.

Shohamy, D., Myers, C. E., Kalanithi, J., & Gluck, M. A. (2008). Basal ganglia and dopamine contributions to probabilistic category learning. *Neuroscience and Biobehavioral Reviews, 32*(2), 219-236.

Siderits, M. (2013). Buddhist Non-Self The No-Owners Manual. In S. Gallagher (Ed.), *The Oxford handbook of the self* (Papaerback edition. ed., pp. 297-315). Oxford: Oxford University Press.

Sitnikova, T., Salisbury, D. F., Kuperberg, G., & Holcomb, P. I. (2002). Electrophysiological insights into language processing in schizophrenia. *Psychophysiology, 39*(6), 851-860.

Skinner, B., F. . (1957/2014). *Verbal behavior*: BF Skinner Foundation.

Smith, L., B., Jones, S., S., & Landau, B. (1992). Count nouns, adjectives, and perceptual properties in children's novel word interpretations. *Developmental Psychology, 28*(2), 273.

Smith, L. B., Jones, S. S., & Landau, B. (1996). Naming in young children: A dumb attentional mechanism? *Cognition, 60*(2), 143-171.

Snijders, T. M., Vosse, T., Kempen, G., Van Berkum, J. J., Petersson, K. M., & Hagoort, P. (2009). Retrieval and unification of syntactic structure in sentence comprehension: an FMRI study using word-category ambiguity. *Cerebral Cortex, 19*(7), 1493-1503.

Speas, P., & Tenny, C. (2003). Configurational properties of point of view roles *Asymmetry in Grammar* (pp. 315-345): John Benjamins.

Stanovich, K. E., & West, R. F. (1983). On priming by a sentence context. *Journal of Experimental Psychology: General, 112*(1), 1.

Stephan, A. (1999). Are animals capable of concepts? *Erkenntnis, 51*(1), 583-596.

Storms, G., De Boeck, P., & Ruts, W. (2000). Prototype and Exemplar-Based Information in Natural Language Categories. *Journal of Memory and Language, 42*(1), 51-73.

Tarski, A. (1944). The Semantic Conception of Truth: and the Foundations of Semantics. *Philosophy and Phenomenological Research, 4*(3), 341-376.

Taylor, J. R. (1995). *Linguistic Categorization: Prototypes in Linguistic Theory*. Oxford: Oxford University Press.

Tenke, C. E., & Kayser, J. (2001). A convenient method for detecting electrolyte bridges in multichannel electroencephalogram and event-related potential recordings. *Clinical Neurophysiology, 112*(3), 545-550.

Thabane, L., Ma, J., Chu, R., Cheng, J., Ismaila, A., Rios, L. P., . . . Goldsmith, C. H. (2010). A tutorial on pilot studies: the what, why and how. *BMC Medical Research Methodology, 10*, 1-1.

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind, 59*(236), 433-460.

Tversky, A., & Kahneman, D. (1975). Judgment under Uncertainty: Heuristics and Biases. *Science, 185*(4157), 1124-1131.

van Berkum, J. J., Zwitserlood, P., Hagoort, P., & Brown, C. M. (2003). When and how do listeners relate a sentence to the wider discourse? Evidence from the N400 effect. *Brain Res Cogn Brain Res, 17*(3), 701-718.

Verheyen, S., Hampton, J. A., & Storms, G. (2010). A probabilistic threshold model: Analyzing semantic categorization data with the Rasch model. *Acta psychologica, 135*(2), 216-225.

Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happe, F., Falkai, P., . . . Zilles, K. (2001). Mind reading: neural mechanisms of theory of mind and self-perspective. *Neuroimage, 14*(Pt 1), 170-181.

Vogeley, K., & Fink, G. R. (2003). Neural correlates of the first-person-perspective. *Trends Cogn Sci, 7*(1), 38-42.

Vosse, T., & Kempen, G. (2000). Syntactic structure assembly in human parsing: a computational model based on competitive inhibition and a lexicalist grammar. *Cognition, 75*(2), 105-143.

Walla, P., Greiner, K., Duregger, C., Deecke, L., & Thurner, S. (2007). Self-awareness and the subconscious effect of personal pronouns on word encoding: a magnetoencephalography (MEG) study. *Neuropsychologia, 45*(4), 796-809.

Wang, K., Cheung, E. F., Gong, Q. Y., & Chan, R. C. (2011). Semantic processing disturbance in patients with schizophrenia: a meta-analysis of the N400 component. *PLoS ONE, 6*(10), e25435.

Ward, J. (2015). *The Student's Guide to Cognitive Neuroscience*. Hove: Psychology Press.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale–Fourth Edition (WAIS–IV)*: San Antonio, TX: The Psychological Corporation.

White, P. A. (2015). The pre-reflective experience of "I" as a continuously existing being: the role of temporal functional binding. *Conscious Cogn, 31*, 98-114.

Wiltschko, M. (2017a). Ergative Constellations in the Structure of Speech Acts. In J. Coon, D. Massam, & L. D. Travis (Eds.), *The Oxford Handbook of Ergativity*: Oxford University Press.

Wiltschko, M. (2017b). Response particles beyond answering. In L. R. Bailey & M. Sheehan (Eds.), *Order and structure in syntax I: Word oder and syntactic structure* (pp. 241-279). Berlin: Language Science Press.

Wittgenstein, L. (1921). *Tractatus Logico-Philosophicus* (D. F. Pears & B. F. McGuinness, Trans.). London: Routledge Classics.

Wittgenstein, L. (1953/2009). *Philosophische Untersuchungen = Philosophical investigations*. Chichester (West Sussex): Wiley-Blackwell.

Wittgenstein, L. (1958/1984). *The Blue and Brown Books.* . Oxford: Blackwell.

Woodman, G. F. (2010). A Brief Introduction to the Use of Event-Related Potentials (ERPs) in Studies of Perception and Attention. *Attention, perception & psychophysics, 72*(8).

Woods, D. L., Kishiyama, M. M., Yund, E. W., Herron, T. J., Edwards, B., Poliva, O., . . . Reed, B. (2011). Improving digit span assessment of short-term verbal memory. *Journal of clinical and experimental neuropsychology, 33*(1), 101-111.

Zadeh, L. A. (1965). Fuzzy sets. *Information and control, 8*(3), 338-353.

Zadeh, L. A. (2015). Fuzzy logic—a personal perspective. *Fuzzy Sets and Systems, 281*, 4-20.

Zimmerer, V. C., Watson, S., Turkington, D., Ferrier, I. N., & Hinzen, W. (2017). Deictic and Propositional Meaning—New Perspectives on Language in Schizophrenia. *Frontiers in Psychiatry, 8*(17).

8 APPENDICES

8.1 The 10-20 system of electrode placement

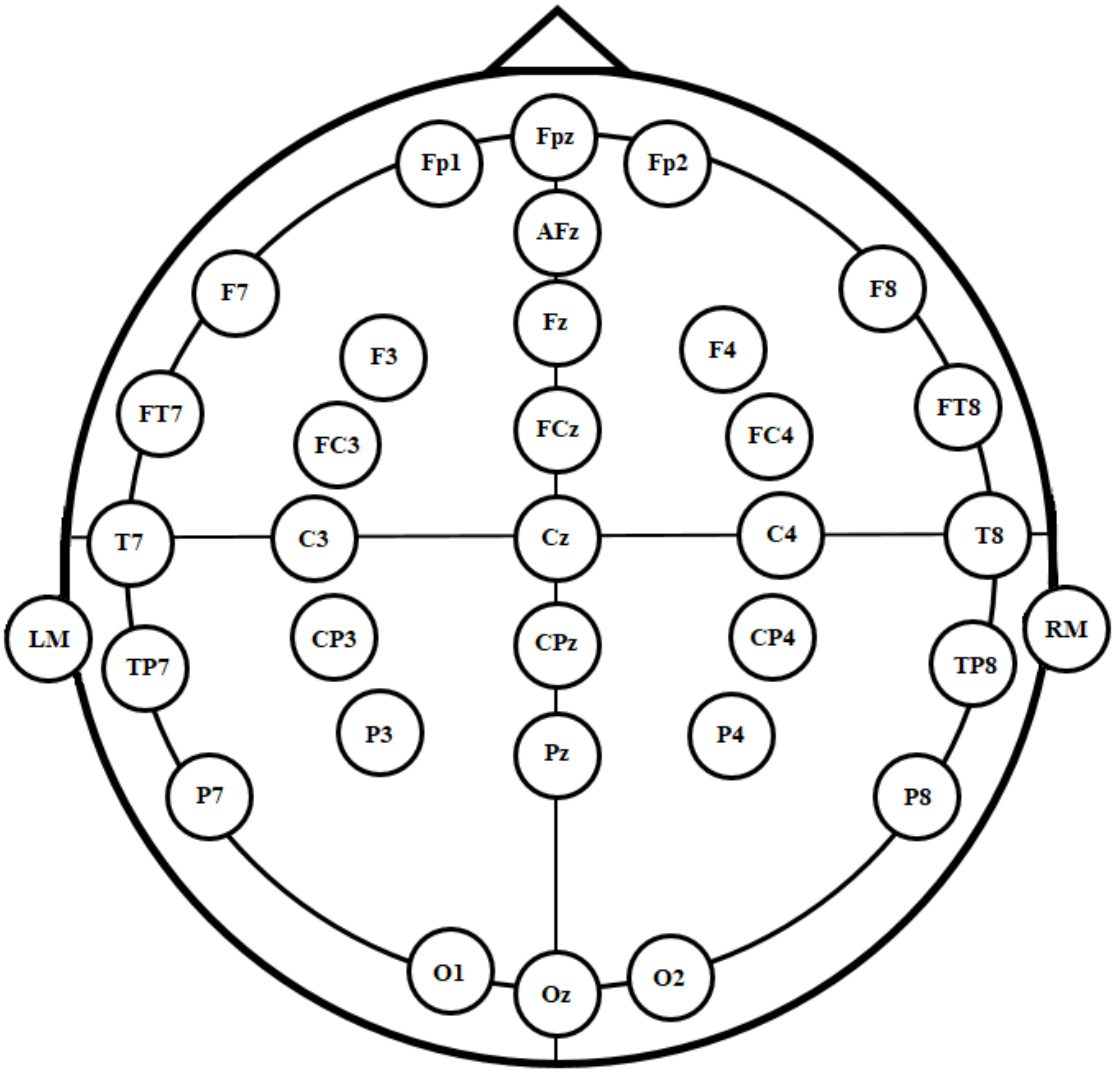

Figure 19: The above 10-20 system of electrode placement was used in both EEG experiments. The letters 'F', 'T', 'C', 'P', and 'O' stand for 'Frontal', 'Temporal', 'Central', 'Parietal' and 'Occipital', respectively. The letter 'z' refers to electrodes on the midline. Even numbers refer to right hemisphere and odd numbers refer to the left hemisphere.

8.2 Additional EEG figures

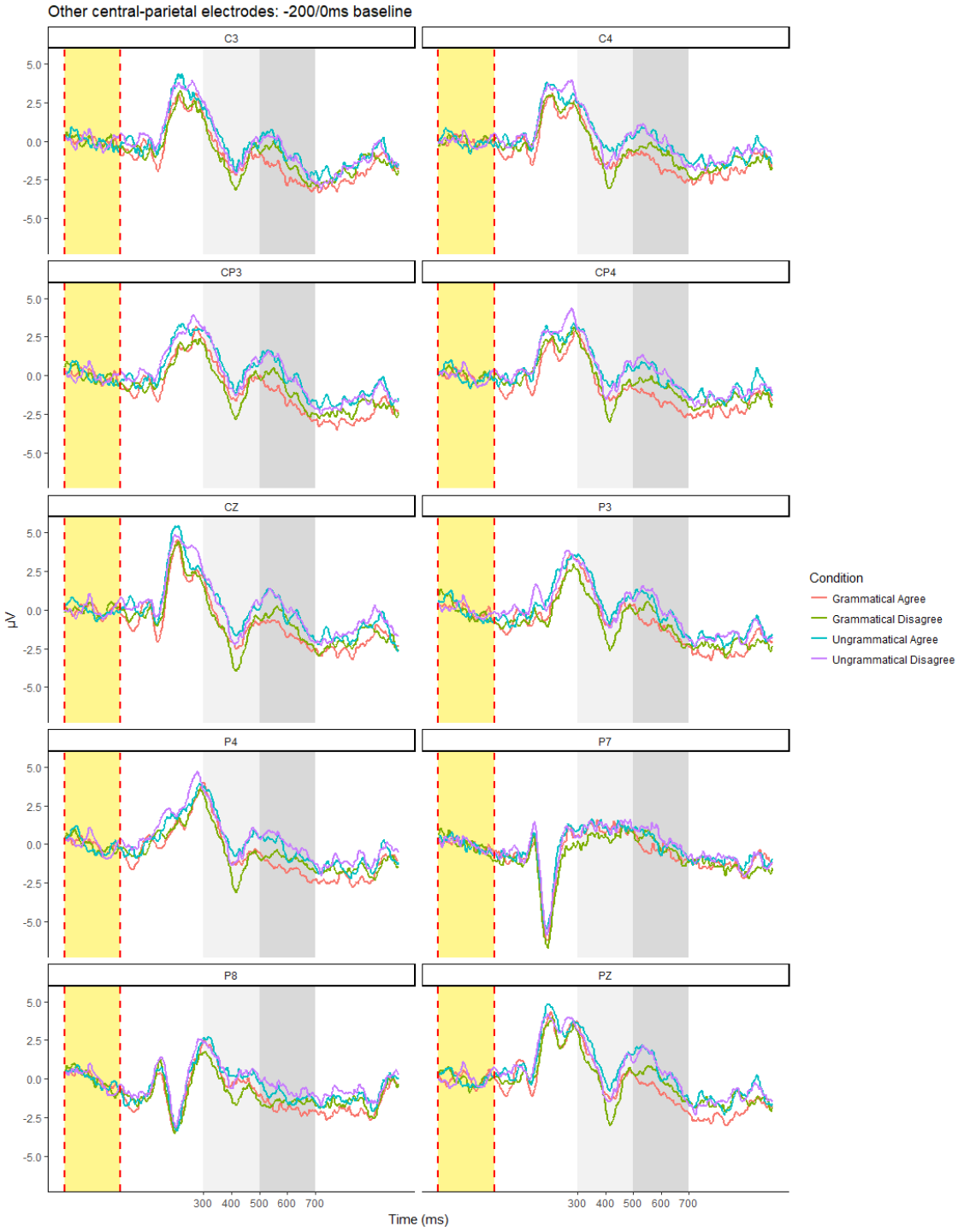

Figure 20: ERP waveforms of all other central-parietal sites for the four experimental conditions. The khaki area represents the baseline (-200ms-0ms) the light grey area represents the N400 window (300ms-500ms) and the dark grey area represents the P600 window (500ms-700ms).

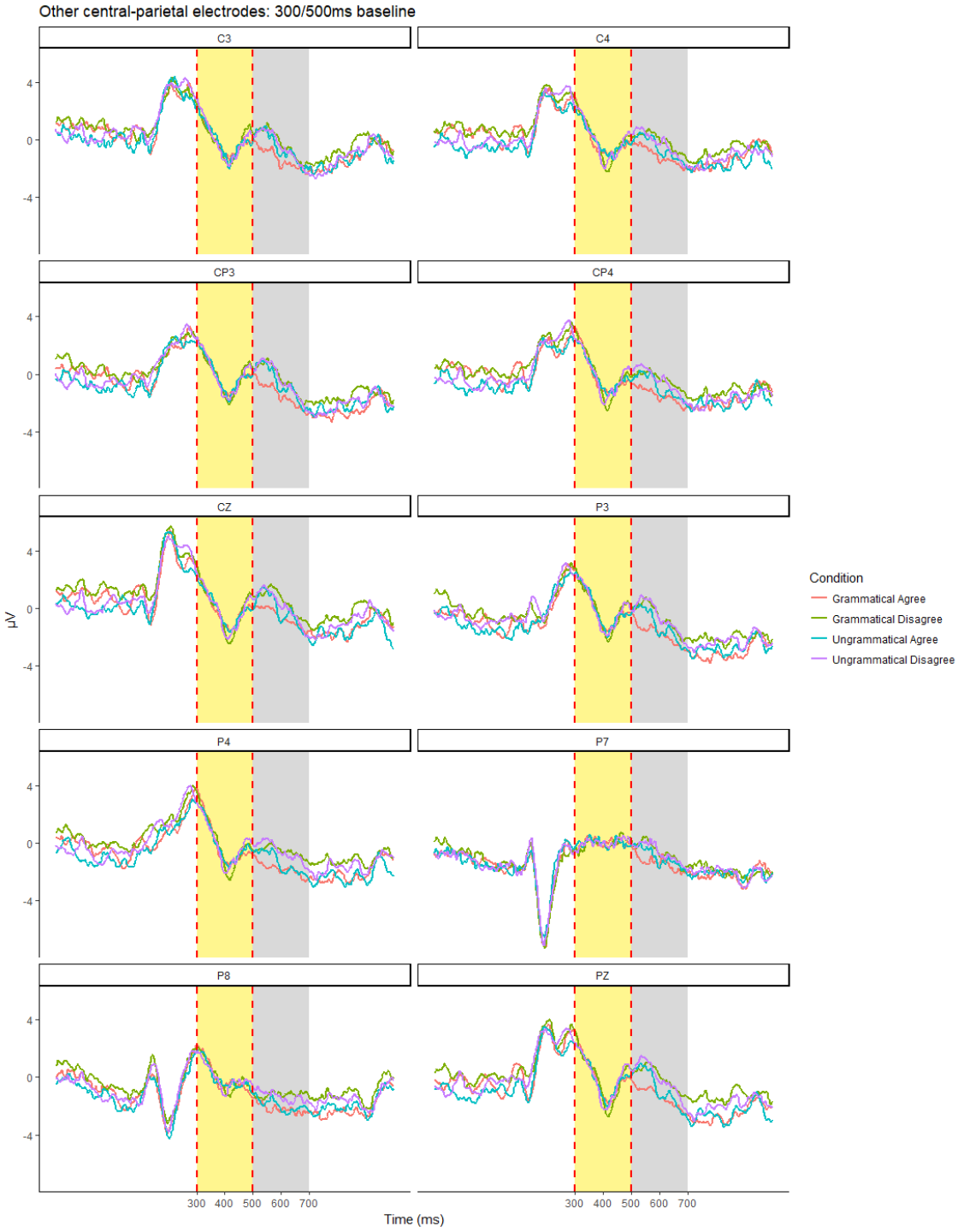

Figure 21: ERP waveforms of all other central-parietal sites for the four experimental conditions. The khaki area represents the baseline (300ms-500ms) and the grey area represents the P600 window (500ms-700ms).

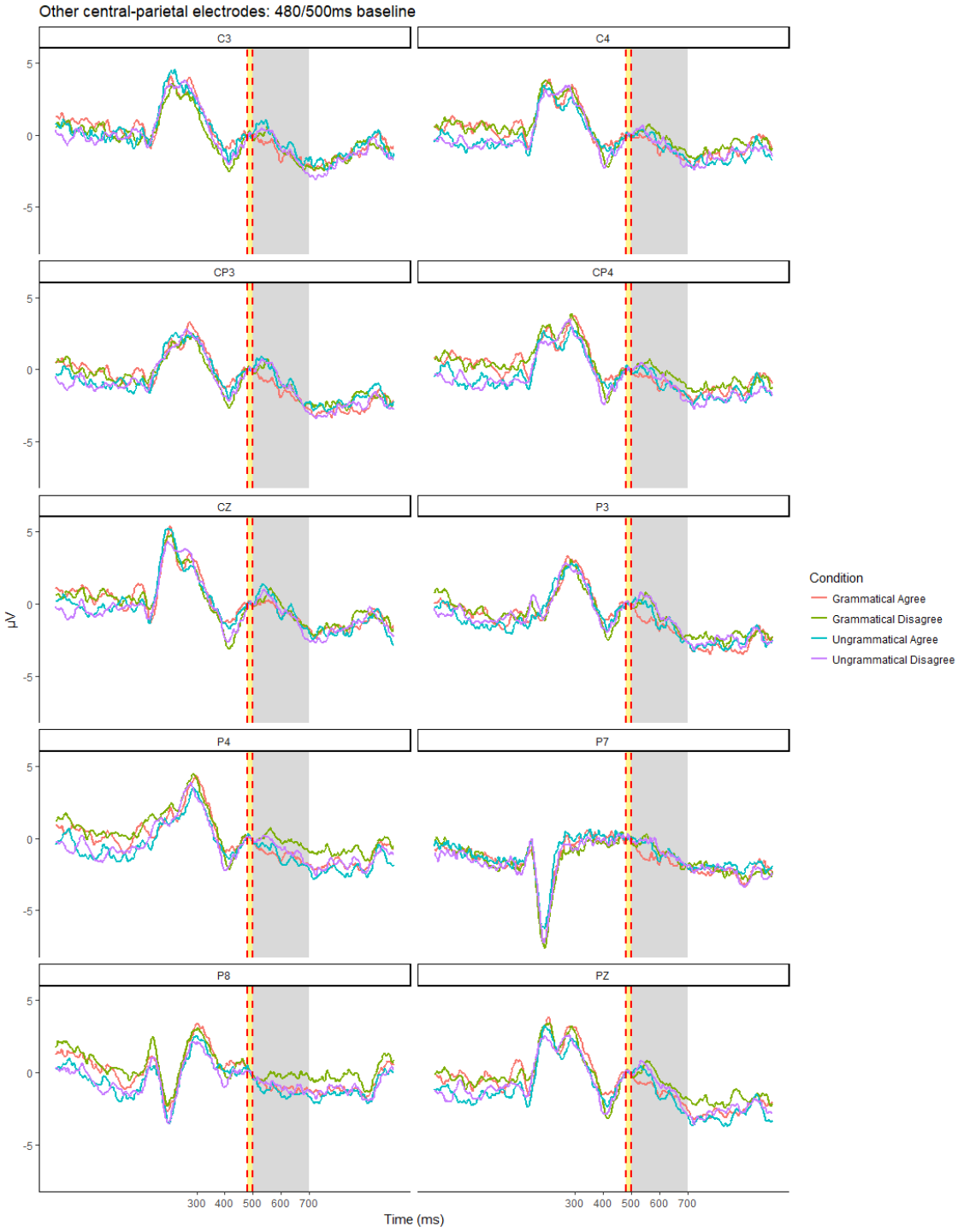

Figure 22: ERP waveforms of all other central-parietal sites for the four experimental conditions. The khaki area represents the baseline (480ms-500ms) and the grey area represents the P600 window (500ms-700ms).

*****