The Meanings of Meaning

Rowan P. Sommers

Thesis for obtaining a “Master of arts” degree in philosophy

Radboud University Nijmegen

Student number: s0849693

Supervisors: Prof. dr. Marc Slors & Prof. dr. Corien Bary

Date: 10-09-2022

Word count: 11078

I hereby declare and assure that I, Rowan P. Sommers, have drafted this thesis independently, that no other sources and/or means other than those mentioned have been used and that the passages of which the text content or meaning originates in other works - including electronic media - have been identified and the sources clearly stated.

Place: Nijmegen, date: 19-09-2022

Abstract

Two interlocutors can (initially) mean different things yet still communicate. Existing theories of meaning are unable to explain such phenomena of semantic heterogeneity because of commitments to strong notions of reference and successful communication. By watering down these notions, the framework of lossy semantics introduced here, hopes to explain semantic heterogeneity. Instead of static, homogeneous word-object relations, lossy semantics sees reference as dynamically updated perceived patterns in our linguistic behaviour; instead of strictly shared meanings, lossy semantics proposes a fail-safe procedure of communication: interlocutors first use preconceived patterns, but when this fails, individuals dynamically establish new patterns based on the perceived behaviour available in the immediate context. This results in more individualistic, pluralist and dynamic notions of reference and communication.

1. Introduction

Quite a few years ago, during the annual family visit to our favourite Chinese all-you-can-eat-buffet, my uncle told me that he used to own marmots when he was a child. He told me quite a bit about them: he had two of them, they were long-haired and often made loud whistling noises. Amazed by his story I asked him whether it had been difficult to acquire them, given that marmots are quite rare in the Netherlands. Slightly annoyed he answered: “Surely you know that you can buy marmots at any pet store.” His answer initially confused me, for I knew of no pet store that sold marmots. Yet it didn't take long before I realised my (and his) mistake - my uncle had never owned any marmots as a child. Sure, he might have had two pets, hairy and whistling, and he must have been calling them “marmots” then too, but still, I was quite confident that the two pets he kept as a child were not marmots, but guinea pigs! Clearly, my uncle somehow believed that guinea pigs were called “marmots”.

Yet, when I told my uncle of his mistake, he would not budge. Abusing the authority of an adult speaking to an ignorant child, he refused my attempt at correcting him, offered a creative taxonomy probably not found in traditional biology textbooks – the short-haired ones might be called “guinea pigs”, but the long-haired ones are surely called “marmots” - and continued talking about the “marmots” he used to own. At the time I lacked the courage (or maybe I was just bright enough to see the futility of the entire situation) to keep correcting his linguistic mistakes, and decided to simply keep quiet and listen to his “marmot” stories. After all, who was I to tell my uncle what kind of creatures his “marmot”-utterances should refer to?

What struck me was that despite his blatant abuse of our common language, I was able to clearly understand what he meant: his “marmot”-utterances, in my mind now too, started to refer to guinea pigs, not marmots. Sentences like “Marmots are the perfect pets!” started to ring perfectly true, despite not knowing anyone who had a marmot for a pet.

Ever since then, I have experienced many such occurrences – where people misuse a certain term, while still making perfect sense. This paper is about this phenomenon, one that any serious theory of meaning should be able to explain. As we shall see, this is a more complicated matter than it at first appears and one which is currently not adequately explained by existing philosophical theories of meaning.

Philosophers of language typically want their theories of meaning to explain i) how language refers to real objects in the world, and ii) how we can communicate with one another. To satisfy these goals, many philosophers have chosen to reject psychologism and now believe that meaning must be understood as a social and publicly shared phenomenon, defined at the level of the language community. Although intuitively reasonable goals, I will argue that current theories of meaning that try to jointly satisfy these two explanatory goals will have difficulty either explaining reference or explaining how I am able to understand my uncle’s “marmots”.

To be precise, my uncle’s “marmot” utterances raises four questions. How can my uncle’s use of the word “marmot” be different from most people’s “marmot” usage? How was I nevertheless able to understand my uncle’s “marmot” utterance? How could the truth values and inferences related to my

uncle's "marmot"-utterances be affected, that is, how could "marmot" genuinely seem to refer to guinea pigs? And finally, how did my understanding of "marmot" rapidly change in order to understand my uncle?

The above example shows that despite the intuitive existence of a socially determined meaning of a word, individuals somehow possess the semantic autonomy that lets their words refer to things other than what they should refer to, as dictated by the reigning social norms. Moreover, it proves the existence of semantic heterogeneity in our societies: people can differ in what they believe words should refer to (Davidson, 1986). Finally, it shows that semantic autonomy is not just the cause of, but also the solution to miscommunication: we can overcome differences in meaning by dynamically altering and aligning our referential judgements to those of others.

In this paper I will attempt to answer the following question: what would a theory of meaning look like that can explain the semantic autonomy and heterogeneity of individuals? I shall start by reviewing why existing theories of meaning have so far failed to reach their explanatory goals: either they fail to fix reference to the world, or they do not have the means necessary to explain semantic autonomy and heterogeneity. I will then propose a new theory of meaning, called lossy semantics, which tries to capture the dynamically changing meanings of heterogeneous groups of individuals, while still making sense of reference to the world and successful communication. Finally, we shall apply this theory to my uncle's "marmots" in an attempt to answer the four questions above.

2. Existing theories of meaning and reference

This section describes three theories of meaning that each attempt to characterise meaning in different ways. They share two common goals: the goal of explaining how we can successfully refer to external objects, and a desire to explain everyday communication, yet none provide a satisfactory account of both.

2.1. Fregeanism

The first theory is often taken as the start of philosophy of language proper. Fregeanism, also called descriptivism, started with Frege (1879) and was, among others, further developed by Russell (1905), Strawson (1950) and Searle (1958) into various forms. Roughly, it consists of three claims:

1. A term refers to an object, when there is a description that uniquely identifies an object.
2. There is a distinction between the sense (the description) and the reference of a term, with the meaning of a term being its sense, and the object that this meaning pinpoints its reference.
3. Sense determines reference. Fregeanism is thus an indirect theory of reference: when a speaker utters a name, a hearer first needs to grasp the correct "sense" before they can grasp the referent of the spoken name.

For instance, the name "Donald Trump" can refer to Donald Trump through the description "the 45th president of the U.S.". Since there is only one individual picked out by that description, through it, speakers of the language can identify to whom "Donald Trump" refers.

In the latter half of the 20th century, several problems with Fregeanism were discovered. Saul Kripke (1980) argued that - in the case of proper names -

Fregean senses or descriptions are unable to uniquely determine reference. The vagueness and arbitrariness with which people typically conceive of persons and objects when they want words to refer to things in the world renders the success of reference quite mysterious. If all we have in mind is fuzzy and arbitrary descriptions of Donald Trump, how can our use of "Donald Trump" pick *him* out rather than some other person who equally fits the particular descriptive content we had in mind during the utterance?

Hilary Putnam (1975) provided a second line of argumentation against Fregeanism: the Twin Earth argument. In this thought experiment, Putnam asks us to imagine a planet which is molecule for molecule identical to ours, except for its water molecules. Instead of water consisting of H_2O molecules, water on this other planet consists of XYZ molecules. You would not be able to immediately tell that from the macroscopic properties of XYZ-water however: people on Twin Earth drink it, swim and shower in it, and call it "water" in exactly the same way as people on Earth do. The argument then goes as follows: a person on Earth and his doppelgänger on Twin Earth are physically and psychologically identical, yet when they use the word "water", and think "water"-thoughts, their words and their beliefs are about different things. A person on Earth talks and thinks about H_2O, while his twin talks and thinks about XYZ. A person's physical and mental state alone therefore provides insufficient information for determining the referents of his words and contents of his thoughts. In addition to the speaker's mental state, one also needs to know the speaker's actual physical state in order to know what he meant.

Later work by Tyler Burge (1979) and Donald Davidson (1987) has subsequently purported to show using Twin Earth-like arguments that we need to add even more than just the actual physical environment: the social environment and the history of the speaker need to be taken into account as well if we are to determine referential meaning. Consequently, if we follow the Fregeans in their assumption that to grasp the sense of a word is to be in a certain psychological state, then Twin Earth-like arguments show that grasping a sense is insufficient for uniquely identifying a referent.

These arguments convinced many philosophers that Fregeanism fails to explain reference to the world, in part because it construes meaning as dependent on the mental states of individuals.[1] These individuals have wrong or at best partially correct ideas about the objects they try to refer to. Hence, subsequent philosophers attempted to ground reference by construing meaning as a publicly shared phenomenon. Though Fregeanism has many real problems with determining reference, we will next see that subsequent theories tended to overcorrect the mistakes of the Fregeans.

2.2. Millianism

Instead of trying to determine reference through descriptions, the Millians attempted several different approaches to fix reference. I shall discuss three versions of Millianism. They all share the belief that reference is fixed directly, without intermediary descriptions or idiosyncratic beliefs of individuals. They differ in the external factors they think are relevant for determining reference instead.

[1] Some Fregeans, Strawson (2002) for instance, have proposed more social versions of descriptivism.

The first and most prominent form of externalism is causal-historical externalism (Donnellan, 1970; Kripke, 1980): instead of being fixed by uniquely identifying descriptions, references of a name are fixed by a causal-historical chain that starts with an initial baptism of the object/natural kind and ends the moment the word is uttered by a speaker. Speakers that use a particular name do not need to have the appropriate description in mind. The speaker needs only to be part of the appropriate causal chain in order for their words to refer to a person or natural kind.

Another variety of externalism, called anti-individualism, puts an emphasis on the social determinants of meaning: the meanings of words are not determined by individuals, but by groups of people (Burge, 1979). Therefore, one cannot determine the meanings of an individual's words by only looking at their internal state. Instead, the social context must also be considered to determine reference.

Finally, Putnam placed a larger emphasis on the division of linguistic labour between experts and ordinary people (Putnam, 1975): Oftentimes, when people use natural kind terms like "water", they are unaware of the scientific concepts involved in these natural kinds that would be necessary to reliably identify the natural kind. Yet, this does not prevent them from successfully referring to them. People can do so because they rely on others in their linguistic community, usually experts, to properly fix reference.

There are several problems with Millian theories. I will limit this discussion to those problems caused by the Millian dependence on social context to provide

meaning.[2] Recall that in the example of my uncle, the dominant social conventions and norms were broken; he invented his own expert opinion that certainly deviated from that of real experts; and while there was a causal-historical chain connecting "marmot" to guinea pigs - a strange off-shoot from the main causal chain[3] - I had not been part of it prior to the conversation.

Thus, social norms, expert opinion and causal history were all working against my uncle: his "marmot"-utterances should never have been able to refer to anything other than marmots. And yet, they did. The social factors of Millians were simply insufficient to fix the reference of "marmot" to guinea pigs and were thus unable to explain our eventual communicative success. Instead, I was able to glean from what little hints my uncle gave me, the type of creature he was actually talking about, thereby somehow overcoming my preference for the meaning dominant in my own social context. Therefore, the social context of the word "marmot" was neither sufficient, nor was it necessary for me to determine my uncle's reference. This directly contradicts one of the main tenets of Millianism: that the social context is required for fixing the meaning of a term (Burge, 1979).

Some philosophers might be inclined to interpret my uncle's "marmot" utterance as referring to marmots anyway. After all, according to both public and expert opinion, as well as the main causal-historical chain, my uncle is simply wrong for thinking that it refers to guinea pigs. And maybe he himself might even

---

[2] Other prominent problems with Millianism are the qua-problem (Devitt & Sterelny, 1999), the problem posed by non-existent objects (Bird & Tobin, 2022), the question of whether or not natural kinds truly exist (Dupre, 1981), and semantic change (Evans & Altham, 1973).

[3] We have attempted to trace the causal-historical chain of my uncle's mistake first to a Dutch tv-show popular in my uncle's youth. In this show, there was a section called the "marmotten race", which involved guinea pigs racing each other through a maze. An even earlier occurrence of "marmot" referring to guinea pigs was in a Dutch 1942 children's novel called "Jeroen en de Zilveren Sleutel".

be convinced of this - if someone with a greater social status than me would have explained the situation to him.

However, this response does entail a restriction of the explanatory scope of millianist accounts: only (mis)communication between people who both follow the same social norms can be explained by such theories. If “marmot” meant marmots, regardless of what my uncle believed, then such a theory does not explain the (mis)communication that clearly occurred between me and my uncle. And importantly, anyone who diverges from the mainstream meaning, like my uncle, cannot mean what they (think they) mean under the yoke of these social theories of meaning. Semantic autonomy would go out of the window.

Millians were of course aware that some form of semantic heterogeneity existed - in fact, the fickleness of the referential judgements of individuals was what made them reject Fregean descriptions in the first place (Kripke, 1980; Putnam, 1975). After all, if every individual had different and misguided descriptions in their minds, and these descriptions determined what their words referred to, then how could these different descriptions ever successfully refer to the same objects in the world? It was not unreasonable for them to prefer their top-down semantic policies over what is admittedly semantic anarchy. Yet, by opting for causal-historical determinism, for tyranny of the masses, or for semantic technocracy, the Millians have inadvertently left no room for the semantic autonomy that people like my uncle evidently have. And it is in large part because of this autonomy at the level of the individual that social interpersonal dynamics are more complicated than the simple, homogeneous and static picture the Millians often sketch of our linguistic interactions and language

communities. As a consequence, Millians cannot properly explain the (mis)communication and the dynamic nature of the semantically heterogeneous folk.

2.3. Two-dimensionalism

The third theory is the currently popular two-dimensionalist framework. Two-dimensionalism is an attempt at synthesising Fregeanism and Millianism. It combines the Fregean and Millian ideas as two separate dimensions within the framework of a possible world semantics, allowing the two-dimensionalist to capture both Fregean and Millian intuitions (Chalmers, 2004, 2006; Jackson, 1998, 2004). Thoroughly discussing the framework, with its merits and its faults will go beyond the goals of this paper, so this overly brief discussion will have to suffice (though see the Appendix for an expanded version).

Two-dimensionalism fixes many of the important problems that were plaguing both Fregeanism and Millianism. I wholeheartedly agree with some of the foundational desires and principles behind the framework - its semantic pluralism for instance. I also admire the technical, nearly mathematical, tour de force of the two-dimensionalist framework. Yet, I cannot help but view this particular product of conceptual engineering as a theoretical Frankenstein. The glueing together of the thesis of Fregeanism and the antithesis of Millianism, has not lead to genuine synthesis: the resulting two-dimensional matrix might have the appearance of unity, but in the way they combined the Fregean and Millian ideas together, two-dimensionalists not only combined the best of both worlds, they also combined the worst (see Chalmers, 2006 or Schroeter, 2021 for nice overviews of its problems).

Moreover, even two dimensions proved insufficient: in order to properly deal with impossible beliefs that may be held by individuals, it introduces a notion of reference where words refer to yet another, third dimension containing impossible worlds that defy the laws of both physics and logic (Bjerring, 2013; Chalmers, 2011). Even if we indulge them with this metaphysical extravagance, the theory remains neither elegant nor parsimonious. Of course, having an ugly formalism does not necessarily imply falsity, yet it does make a good reason for trying to find a better theory of reference elsewhere.

3. New alternative: Lossy semantics

So far, previous theories of meaning have not adequately explained referential meaning. Throughout the various attempts, we can observe a pattern, a recurring problem: an insurmountable tension in the explanatory desires of philosophers. The root of the problem is that philosophers of language on the one hand want that their theories of meaning are able to objectively determine reference to objects in the world as they really are, and on the other want them to explain communicative behaviour of individuals who often fail to see the world as it truly is. These desires are not easy to jointly satisfy. Fregean theories of meaning have given priority to the second desire, at the cost of the first. Millian theories focused on objectively determining reference, at the cost of explaining the behaviour of individuals like my uncle. Two-dimensionalism, finally, struggled with the same dilemma, and was only capable of solving it through the ad-hoc introduction of an impossible third dimension, burdening its framework with ever more complexity.

In sum, the history of philosophy of language gives no indication that these two desires - in their current form - are jointly satisfiable in a satisfying manner.

To go beyond these two conflicting desires I suggest the following alternative strategy: we relax both the assumption of reference to the world and the assumption of successful communication as a one-to-one mapping between speaker meaning and hearer meaning. Instead, we stick closer to actual linguistic behaviour. By constructing a new formalism based on our linguistic behaviour, we shall make room to loosely talk of reference and successful communication. In turn this will lead to a theory capable of explaining the four desiderata of my uncle's "marmot".

The assumption of reference to the world can be relaxed by proposing a new indirect theory of reference, grounded in the subjectivity of the individual. In the framework which we shall dub lossy semantics, referential meaning can be made more or less definite and explicit. But it ends up as necessarily time-relative, pluralist, and impossible to objectively observe in its totality. To get to lossy semantics however, we will have to proceed in steps and slowly deconstruct the prevailing ideas surrounding reference. In this process, we cannot avoid making some simplifying assumptions about both reference and intentionality that will slowly be dissolved as we proceed with the deconstruction process.

First, we will set up a fine-grained description of an individual's idiolect: a complete set of facts about an individual's referential behaviour. By concatenating multiple idiolects we form a fine-grained dialect. From this "ground truth", we will then try to extract coarse-grained individual meanings and even coarser-grained group-level meanings that should help us to identify reference. These are still more or less objective. Finally, we shall compare these ontological descriptions of meaning to our epistemological reality. This exercise will serve to

show, first, what a more individualised notion of meaning is, and second, why the type of referential meaning that we actually possess is inherently subjective and partial.

One final note, I will limit myself to a discussion of natural kinds only, but I fully expect that the proposed framework generalises to all content words, and can perhaps even explain function words like “the” and “only”.

3.1. Idiolect and dialect tables

How should we determine someone’s idiolect? How do we typically tell what another person means? In practice, we infer it from their behaviour, linguistic or otherwise. Let us look at an example that might capture the phenomenon.

Imagine a young boy named Billy helping his father paint the front of their house. Suppose further that his father gave young Billy the all-important mission to fetch him a glass of water - to quench his thirst. Imagine next that Billy quickly ran to the kitchen to find some water and that arriving there, finds a full glass sitting on the kitchen counter already. After taking out the paintbrush, and pinching his nose in surprise at the smell, he brings it to his father and says: “Dad, wataa!” Now let us ask ourselves: what does Billy mean by “wataa”? From his immediate behaviour it seems obvious that he means the stuff he is holding in his glass. But it is possible that the type of stuff he is holding is not the only thing he would call “wataa”.

A single situation is thus insufficient to fully determine with certainty and precision the full range of substances which Billy takes to be denoted by “wataa”. After all, there could be other substances than the one he found in the kitchen that he could have brought to his father. We should therefore observe his “wataa”

behaviour in other situations as well, to see which objects are denoted by it. This would bring with it the risk of Billy changing his meaning of “wataa” halfway, but it would still be informative.

If we were all-knowing gods, then we could be more systematic. We would not only observe the totality of Billy’s behaviour in the actual world, but we could do additional behavioural experiments in alternate possible worlds. For instance, we could place young Billy in a possible world with a different liquid on the kitchen counter, so we could see which of these he would bring to his father. Doing that for all possible liquids - or better, all possible objects - would give us a good sense of his “wataa” beliefs. Alternatively, we could simply ask him directly whether he would consider a certain liquid “wataa” or not. Then repeat the same experiment for all possible liquids (or objects). Finally, we can ask him general questions about what “wataa” is and see how he responds. With an infinite number of possible worlds, there are an infinite number of experiments that could be performed to more fully nail down what Billy thinks “wataa” means.

We could then observe his behaviour in response to other words and other sentences. We could present him with non-words, pseudo-words, jabberwocky sentences and completely nonsensical sentences to see how he reacts. We could even observe Billy’s behaviour to all of the possible gestures we use while talking, or to all signs and non-signs that exist in the various sign languages. Finally, to be extra thorough, we could do these experiments for all languages - all possible languages - all possible sound waves even. Essentially, to get a complete picture, not just of Billy’s use of the term “wataa”, but of his entire idiolect, we would observe his behaviour in all possible scenarios or worlds.

Should we be forgetful gods, we would document our observations in an extended table similar to Table 2. This table would then contain all that there is to know about Billy's idiolect without direct access to Billy's inner state. Let us call this type of table an idiolect table. We could of course further generalise this idea to the social domain by concatenating the tables of multiple individuals together. Let us call these concatenated tables dialect tables.[4]

[4] Note that in the creation of these tables we already assume a primitive notion of reference and intentionality that we have not yet defined. Filling in the entries of these tables entails all kinds of assumptions about the nature of natural kinds, about which types of behaviour can inform us towards which specific objects our linguistic actions are directed, and so on. In sections 3.3 and 3.7, I will devote another footnote each to how a lossy semantics definition of reference more accurately involves this table.

Table 2

Idiolect Table of Billy

Entry	Possible world	Behavioural outcome
1	Dad: “Could you fetch me a glass of water, Billy?” A glass of white spirit sits on the kitchen counter.	Billy takes the glass of white spirit to his father and says “Dad, wataa!”
2	Dad: “Could you fetch me a glass of water, Billy?” A glass of pure H_2O sits on the kitchen counter.	Billy takes the glass of pure H_2O to his father and says “Dad, wataa!”
3	Dad: “Could you fetch me a glass of water, Billy?” A glass of pure XYZ sits on the kitchen counter.	Billy takes the glass of pure XYZ to his father and says “Dad, wataa!”
4	Dad: “Could you fetch me a glass of water, Billy?” A glass of tap water (H_2O + assortment of minerals) sits on the kitchen counter.	Billy takes the glass of tap water to his father and says “Dad, wataa!”
5	Mom: “Could you fetch me a glass of water, Billy?” A glass of tap water (H_2O + assortment of minerals) sits on the kitchen counter.	Billy takes the glass of tap water to his mother and says “Mom, wataa!”
6	Dad: “Could you get me some lemonade, Billy?” A glass of white spirit sits on the kitchen counter.	Billy makes some lemonade with the white spirit, takes it to his father and says: “Dad, lenade!”
7	Dad: “Could you get me a beer, Billy?” A glass of white spirit sits on the kitchen counter.	Billy brings a beer from the fridge to his father and says “Dad, bear!”
⋮	⋮	⋮
∞	Dad: “Slab!”	Billy gives his father a slab

3.2. Reference is time-relative

So far, we have discussed a variation of a behaviourist usage-based account of meaning.[5] We might wonder then, where is reference in all of this? How do we get from the behavioural dispositions of Table 2 to reference to the world? What one might want to say is that in Billy's idiolect, "wataa" means the set of objects that he would call "wataa", those that he would bring to his father when asked. More generally, in Billy's idiolect, the reference of an utterance U is the set of objects R that Billy would behave appropriately towards across all possible worlds.

There are four complications that will entail further changes to this working definition. First, both words and people change meaning over time. Words have no static meanings. Instead, their meanings dynamically shift as a function of time. Common examples of words whose meanings radically changed are "guy", "nice", and "gay": "guy" used to be the proper name "Guy Fawkes"; "nice" meant foolish; and "gay" originally meant joyous or bright (Oxford University Press, 1989c, 1989b, 1989a). These semantic changes are grounded in the fact that individuals - especially young children - continuously change their minds about what they believe words to mean (Bowern, 2019; Hock & Joseph, 2019).

The existence of these changes poses a problem for our working definition of reference, as contradictions can and will occur if we create idiolect and dialect tables containing multiple time points. For instance in the "wataa" example, young Billy did not know at first how to distinguish between water and white spirit and thus decided to bring his father the latter, rather than the first. After his

[5] As we shall see, the resulting notion of reference proposed by lossy semantics will not depend very strongly on behaviourism. There will be sufficient room for mentalist influences on reference later.

father corrects his mistake and teaches him that the smell distinguishes water from white spirit, Billy might successfully retrieve the glass of water his father required. Thus, at time t_1, Billy thinks "wataa" refers to white spirit too, whereas at t_2, he has learned and consequently adapted his behaviour: in his new idiolect, "wataa" no longer refers to white spirit. As a result of this behavioural change, Billy's idiolect tables at the respective time points will contradict each other, thereby making Billy's overall idiolect inconsistent.

The solution to this is simple: we make reference time-relative. Instead of believing that a single referential relationship can be formulated that can cover all time points simultaneously and still be consistent, we shall assume that reference is defined only relative to a particular point in time.[6] Consequently, in the idiolect of a person P at time t, the reference of an utterance U is the set of objects that P at time t would behave appropriately towards across all possible worlds.[7] For the example with Billy this implies that his extension of "wataa" changes as he receives feedback from his father - as it should.

3.3. Reference is a real pattern

[6] Some might propose to take the time that the utterance was spoken as the true time, but this is not ideal. The same utterance can be interpreted far after the fact, from memory. And we might have a different interpretation of the same sentences at later times in our life. Others might propose that the speaker's meaning (at the time of utterance) determines the true meaning. But in a semantically heterogeneous world, this is not ideal either, as different people might interpret the same utterance in a variety of ways. Instead, I view meaning akin to the meaning of a work of art: everyone interprets it in his or her own way, and while the artists' own intended meaning is special in a way, it is not necessarily better than the others.

[7] Note that this solution circumvents the usual philosophical objections to idiolects based on Wittgenstein's private language argument (Wittgenstein, 2010): no appeals to correctly remembered associations between private sensations are needed. Consistent potential behaviour in a single time slice is sufficient for meaning.

The second problem is the following: our idiolect and dialect tables are filled to the brim with information about our complex semantic behaviour. These tables should provide all the information we need to determine precisely what one means with a given word in all possible situations. Unfortunately, while our behaviour is mostly consistent, this is not always so. Whether the result of mistakes, or because we genuinely disagree about what to call a given object, our semantic behaviour is frequently inconsistent, even within the same time period. This is especially so in dialect tables containing multiple individuals, like the one containing me and my uncle. If we want to determine the reference of a term, which is to find a rule or pattern that maps an utterance to a particular set of objects, then we will need to deal with these inconsistencies.

Suppose that we compare the idiolects of two ordinary English speaking individuals and somehow decide on their extensions of the term “American elm”. Delineating elm species is notoriously difficult. Even if these two individuals were proper experts, there will likely be differences between the two sets of objects identified as “American elm”. Yet, despite these differences, there will most certainly be a great deal of overlap too.

Suppose we then try to formulate a rule that captures these commonalities - we could take the intersection of the two extensions, or we could perhaps find some other pattern in the two sets of objects - perhaps some description like “large deciduous trees from North America”. Alternatively, we could try to specify it in terms of sensory impressions, for instance the specific shape of the leaf of this particular species; or maybe we discover that some percentage of the nucleotides in the DNA of the identified trees overlaps, making that a nice rule of thumb for

determining reference; Perhaps there is a clear causal chain going from a single baptism to the usage of others. Any of these rules shall likely predict a substantial amount of behaviour and can help us decide what "American elm" means in this duo's dialect. Nevertheless, no matter what rule or pattern we devise, the differences in behaviour found in the two idiolect tables are such that there will always be some behaviours that are not correctly predicted by the rule or pattern.

The fact that a pattern or rule such as reference can not perfectly predict our behaviour entails that some entries of the table - those containing behaviours that contradict the rule - will have to be ignored or de-emphasised. The problem with this is that ignoring some entries does incur a cost. By ignoring particular entries, some information contained in the table will get lost. If we decide that the word "marmot" does not refer to guinea pigs, then the entries where my uncle calls a guinea pig a "marmot" cannot be explained.[8] At the same time, simple rules are not without use. While a given referential rule will not be able to explain and predict every single one of our (often erratic) behaviours, it is more than capable of explaining a large subset of them.

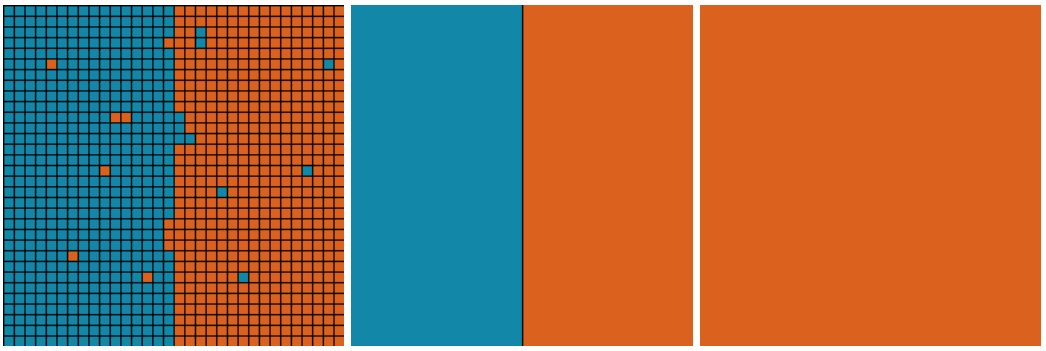
The above characterization of reference is reminiscent of Dennett's notion of real patterns (Dennett, 1991). According to Dennett (1991, p. 34), "[a] pattern exists in some data - is real - if there is a description of the data that is more efficient than the bit map, whether or not anyone can concoct it." Dennett's real patterns are a compromise between on the one hand the realisation that the concepts and patterns we distil are oftentimes imperfect - they do not neatly map

[8] Whether there exists a perfect pattern that neatly explains all of our behaviour is ultimately an empirical question that awaits further developments in cognitive science, but for now, we have no contemporary scientific theory capable of this, and, in all honesty, doubt we will ever attain it. I will therefore assume only the full idiolect and dialect tables will be able to correctly predict our behaviour.

onto the bit map of the physical world and are therefore not exactly real. On the other hand there is the evident truth that our coarse-grained patterns are typically good enough to allow us to survive and understand one another - thus containing a grain of truth. This seems similar to reference, and Dennett's lessons could therefore be of use to us.

Figure 1

An Example of Lossy Compression.



Note. On the left we see a 32x32 pixel bitmap, where each pixel can have one of two colours. In the middle, the same bitmap is lossily compressed into a form that more efficiently describes and predicts what colour each pixel is. Note, that some pixels will be falsely predicted, resulting in some errors. Finally, on the right, an overly compressed depiction of the bitmap is presented that will predict pixel colours at chance level.

The similarities between reference and Dennettian real patterns are striking. For one, real patterns are often lossy compressions. Lossy compression is an information theoretic notion often used in computer science. An example of lossy compression is when you resize an image from a high resolution to a lower resolution. This typically results in a loss of information, as some pixels that might have a slightly different colour are merged together into a single cluster of pixels with a colour that best matches the average colour of the old pixels (see Figure 1). But if it is not compressed too much, the image still contains the

relevant information that was supposed to be conveyed: the overall pattern remains.

Likewise, Dennettian real patterns do not perfectly explain and predict our targets. Not all of the information contained within the original bitmap is retained - similar to how our simple referential rules cannot predict and explain every entry found within the meaning tables. Yet, despite this loss of information, real patterns are still capable of predicting things with a probability greater than chance level - so too does our referential rule.

Finally, Dennett shows that there are often multiple possible descriptions of any given bitmap, and thus that there is more than one real pattern - without us being able to choose between them on non-arbitrary grounds. This lesson will become important later.

If we take this analogy seriously, then it should be clear that the idiolect or dialect tables can serve as our meaning bitmaps.[9] But what exactly are the real patterns that efficiently describe the information therein? After all, there are many ways to compress the meaning tables. We will answer this question later when we discuss the third problem in Section 3.4. For now it will be sufficient to know that there are real patterns in our referential behaviour, that is, there are real patterns in the ways we associate words to objects, and that these real patterns (over)simplify our complicated linguistic behaviour - for better or worse.

[9] As mentioned in footnote 4, taking the idiolect and dialect tables as meaning bitmaps is a simplification. In truth, the meaning tables are themselves coarse-grain descriptions of an even finer-grained physical bitmap. Their entries are constructed by taking what Dennett calls the intentional stance and thus distilling real patterns in the true physical bitmap (Dennett, 1987, 1991). Since they too are real patterns, each entry in the idiolect table therefore faces the same considerations as the larger referential real patterns we create from the tables: they are similarly subject to imperfections, multiplicity and arbitrariness.

With that in mind, let us first consider the problem that we saw earlier: that reference is typically conceived of as an overly simplistic relationship between a word and a set of objects within some larger language community. While this understanding of reference allows one to nicely delineate a discrete set of objects, it does come at the cost of a proper explanation of our complex linguistic behaviour. Especially the behaviours of those who deviate from established social norms.

Instead of this one-size-fits-all approach, I propose a fail-safe notion of reference, modelled on my experience conversing with my uncle. This notion goes like this: Whenever we strike up a conversation with someone, we first use the real patterns distilled from the larger dialect table to decide what a person refers to with any given utterance. This will often be good enough. But since these real patterns are imperfect, lossy compressions, they will inevitably fail on some occasions, resulting in behavioural evidence that contradicts our predictions. Whenever this occurs, we resort to plan B: we use the information in the idiolect table of our interlocutor to establish a new real pattern that could explain their idiosyncratic linguistic behaviour. This can act as a new rule that explains what they mean by the word.

More formally:

Plan A: In the dialect of a community C of individuals at time t, the reference of an utterance U is the real pattern that maps U to a set of objects R, best explaining the U-related behaviour contained in the dialect table of C at t.

If our real pattern fails to explain the behaviour of a person P at time t, that is, if plan A fails, we resort to plan B:

Plan B: In the idiolect of a person P at time t, the reference of an utterance U is the real pattern that maps U to a set of objects R, best explaining the U-related behaviour contained in the idiolect table of P at t.

3.4. Reference is pluralist

This is better, but still not without problems. Our redescription of reference as a real pattern leads us to the third complication. Namely, that there are often multiple equally valid patterns that describe the behaviour of an individual. This idea is related to Quine's inscrutability of reference thesis (Quine, 2013), which states that the relationship between word and object will always remain ambiguous. No amount of empirical evidence will let us completely determine what a particular sound refers to. While the thesis is probably posed too strongly, it is true at least that determining reference is a so-called ill-posed problem (Courant & Hilbert, 2013; Hadamard, 1932): given a limited number of observations (of Billy's "wataa" behaviour for instance), there are multiple possible, equally valid solutions or conceptual schemes that would explain his referential behaviour. Note the similarities with Dennett's claim that there are multiple real patterns that can usefully describe the same bitmap.

This can be illustrated by a number of remaining questions. Some of these questions have to do with further inconsistencies and ambiguities in Billy's behaviour. For instance, how do we deal with inconsistencies stemming from perceptual mistakes? If Billy mistakes a glass of lemonade for water - perhaps in a badly lit setting - should that glass of lemonade be part of his extension of "wataa"? If not, how exactly do we demarcate abnormal and normal perception? Or how do we decide which specific behaviours count as appropriately

meaningful? Pointing and fetching particular objects is quite easy to interpret, but how should we deal with more ambiguous behaviours like a slight nod of the head, a subconscious eye movement, or a slip of the tongue? Which of the possible worlds where these behaviours appear should be considered when determining the extensions of words?

A second set of questions relate to inconsistencies that occur between people. For instance, which people should we put together into the same dialect table? Does someone from Scotland speak the same dialect as someone born in the US? Probably not, but where do we draw the line? Is my uncle's idiolect similar enough to mine that we are part of the same dialect? Depending on how we draw the boundaries around different people, we might get different answers to what a word refers to. And if we do decide to include my uncle in the same dialect table, then how do we deal with inconsistencies that crop up because of semantic heterogeneity? Should we treat my uncle's "marmot" as a separate sense of "marmot"? Or should we - as the Millians desire - disregard his "marmot" sense entirely, and stick either to the majority vote, to the causal-historical chain, or to expert opinion? Should we treat all people equally, or should we perhaps give more weight to the semantic opinions of certain expert individuals? And if so, how exactly do we determine this weight?

These questions - and many more - help to point out that in spite of our access to godly amounts of information, major sources of ambiguity remain. When we try to discern a real pattern in our data, we are forced to compare and arbitrate between different conflicting observations. To be sure, there are likely some rational guidelines to be given here. But some arbitrariness seems

unavoidable when determining reference: we want to choose one possible extension or pattern over another, but often lack clear reasons for doing so. Even if we have all of the information about the world, then still there are multiple patterns that can serve as the reference or meaning of a term, many of them performing far better than chance, and hence real, but not necessarily with one better than the others.

The solution to this problem is to make reference relative to the particular definition or set of semi-arbitrary choices that one can make in determining reference.[10] That is, reference is relative to the way one chooses to deal with referential ambiguities, with different real patterns as a result. As the above list of questions shows, there are many choices that one needs to make. Providing a full list will be far, far beyond the scope of this paper, but we can summarise a great many of them with a couple of important dimensions. In so doing, we will at the same time illustrate what the compressed referential real patterns truly are.

First, in order to compress the information in the meaning tables,[11] one needs to make decisions about which objects and kinds in the entries of those tables are sufficiently similar. Using these similarity judgements, we can determine whether or not the entries of multiple objects/kinds can be usefully merged. In other words, to determine reference, one needs to have an ontology of objects and kinds. For instance, if you are an externalist and believe that the differences between the microphysical properties of H_2O and XYZ are large enough that they are genuinely different natural kinds, then the entries in the

[10] Alternatively, we could make reference relative to the particular goal and purpose one has with the interaction. These two formulations would roughly align, for the set of semi-arbitrary choices depends crucially on these goals.

[11] In fact, one cannot even begin to create the meaning tables without making these decisions.

meaning tables about those respective kinds should not be conflated. If however, you are an internalist, who believes that the microphysical differences of H_2O and XYZ are not essential - for instance because most people interacting with the two substances are unaware of any differences - then to you those entries can be treated as the same. Exactly how and the level of granularity with which one draws the boundaries around objects and kinds in nature will determine what words refer to. Different ontologies yield different real patterns.

Secondly and similarly, we need to know which sound waves (or collections of marks on paper) are sufficiently similar to one another. On top of a list of words like a dictionary, we need a way to map each uttered sound wave or written word onto one of those words (or onto a gibberish-category). For example, do “wataa” and “water” constitute different word forms for the same word, or are they substantially different words? Your answer will determine whether you compare the meaning of “wataa” to “water” or whether you treat them as separate words in your lexicon. Different sound-to-word mappings result in different real patterns.

The third dimension is closest to the way Dennett views his real patterns: as lossy compressions of our own behaviours as seen from the intentional stance. The real patterns we extract from the idiolect table of our interlocutor and that we use to predict their behaviour are imperfect, lossy compressions of the physical system constituting our interlocutor. The ways we classify the movements of this physical system into intentional and meaningful behaviours - by taking the intentional stance - will determine what we believe their words to mean. Different behavioural real patterns will create different referential real patterns.

Finally, the fourth dimension consists of the way we delineate our socio-linguistic communities: what dialect a particular person speaks and how to draw the boundaries around different dialects in the first place. As an example, I initially believed that my uncle was speaking the same dialect as I was, but it turned out that he was speaking some other dialect instead, and my interpretation of the meaning of his “marmot” utterance changed as a result. In sum, different socio-linguistic classifications create different real patterns.

Two notes: none of these dimensions act in a vacuum. They interact with one another, hence, a change in one will often affect the others. It is therefore better to consider them together as a single complex space or web. Note further that the four dimensions each seem to have only two distinct levels: a fine-grained bitmap, and a coarse-grained real pattern that simplifies the information in the bitmap. In reality, there are nearly as many levels as there are patterns in the bitmap. Therefore, the differences between levels are more continuous than it may appear from our discussion so far, e.g. we can group people in various levels of granularity: couples, families, neighbourhoods, schools, football clubs, countries, continents, and on other semi-arbitrary categorizations like whether or not one likes fishing, plays the same videogame or is part of the same social media bubble.

By making these semi-arbitrary decisions about similarities and differences between objects, sounds, movements and persons, we can merge different table entries together. By grouping objects together into kinds; sounds together into words, bodily movements into meaningful behaviours, and people together into

communities,[12] complex reality becomes simpler. By relating these kinds, words, behaviours and communities together, we can summarise the more fine grained entries and form a new meaning table. The new entries of that table - the products of the merger - describe the older entries more efficiently, for the reasonable price of a loss of detail. These are the real patterns in the fine-grained meaning tables.

The resulting lossily compressed meaning table describes fuzzily which object/kind is referred to with which sound/word, in conjunction with which movement/action by which person/community. This just is what reference is. And it means that reference is inextricably linked to a set of semi-arbitrary decisions about how exactly the complex information in the meaning tables should be compressed.

3.5. Reference is only partially perceived

The fourth complication is that we are, in fact, *not* gods.

One consequence of this sobering thought is that we do not have access to all the information about one's idiolect or dialect tables: we cannot know with certainty how a given person will behave in the actual world, let alone how they behave in all possible worlds. We can only ever make one observation per time point and thus only see and hear an astonishingly thin slice of the complete idiolect and dialect tables.

Furthermore, we do not store exact copies of our experiences in memory, but rather incrementally update our beliefs about the world based on incoming information. And since reference is time-relative, and both idiolect and dialect

[12] Do not forget that this is a mere sketch of the dimensions and factors in play. One can think of other schemes. For instance, it might be preferred not to group people into communities, but instead classify their utterances into languages or dialects. After all, a single person can learn to speak many different languages. Delineating different languages is similarly difficult as placing boundaries around speech communities however (Haugen, 1966; Langston & Peti-Stantic, 2014).

tables will change over time, the limited number of observations and our limited memory capacity prevents us from tracking all of the changes that occur in all of the individuals that exist in our community. Reference is therefore a moving target that we can never quite hit.

Moreover, we each perceive different slices of the idiolect and dialect tables. I might see someone use “water” to refer to a body of water X, but not to a body of water Y, you might see them refer to Y, but not to X. The waters differentially shape our perception of what “water” means. Finally, different individuals have different pattern extraction methods, different pattern recognition skills and different decision criteria for determining reference. This variety of experiences, extraction methods, recognition skills and semantic decision criteria all contribute to the fact that despite our best efforts at forming identical patterns, the patterns we perceive will inevitably differ from one another. No wonder then that there exists semantic heterogeneity in our use of language. Hence, the patterns that we will extract are undoubtedly significantly worse and varied than the more optimal (yet still varied) patterns the gods are able to form:

Importantly, it is these suboptimal, perceived patterns that we actually use during communication, not the (pluralist) real patterns only gods have access to. These flawed, partial patterns are the ones determining how we will interpret the meaning of someone else’s term, and how we will subsequently behave. And should we form new patterns - even if done in error and based on foolish ideas - these new patterns nevertheless change our behaviour and thus cause real changes in our idiolect tables.

And yet, despite their flawed nature and these differences in perceived patterns, we seem surprisingly capable of communicating with one another. It is undeniable that our perceived patterns often asymptotically converge to the same or to a similar meaning. This convergence is not sufficient to explain all of our successes in communicating however: while the convergence process is still ongoing, semantic differences remain. And importantly, the convergence process is always ongoing.

Perhaps if we were able to lock up the entire planet in the same room and persuade each person to communicate and align their meanings with every other person, maybe then everyone would start talking in exactly the same language. Alas, in the real world this is impossible. Here, our semantic behaviour definitely aligns but only in a local, piecemeal fashion, governed by our ever changing social networks and continuous efforts to change our word meanings. The gaps in our networks are too large, the changes in meaning are too diverse and too fast for any one individual to pick up immediately. Thus, we never quite get to a perfectly homogeneous language.

Instead, we are stuck with a multitude of (similar) languages that are stable enough to work well most of the time, but in which pockets of subcommunities continuously change their semantic behaviour according to local social dynamics (in historical linguistics this is called the wave model (François, 2014; Schmidt, 1872; Schuchardt, 1885)). And this means that between any two persons there is always the possibility of further semantic differences that prevent immediate successful communication.

The existence of inextinguishable semantic differences demands that we are flexible in our interactions with others. Because we may initially start a conversation with markedly different patterns than our interlocutors, we regularly check whether the other person's behaviour corresponds to what our pattern would predict. If this behaviour violates our predictions, as it often does, we move to plan B. We try to discern a new pattern in their idiosyncratic behaviour.

We could then initiate a linguistic negotiation (Ludlow, 2014): either we try to correct their mistake, or we defer to them and adapt our own behaviour to theirs (Lassiter, 2008). If initial negotiations fail, we might resort to third-parties like dictionaries or experts, or we might agree to disagree and keep both senses of the word. No matter how - or even whether[13] - it is achieved, the goal is renewed semantic alignment (Pickering & Garrod, 2004, 2006). If we are happy with the new pattern, we might choose to update our old pattern, remembering it for future conversations. If enough people start using the new pattern, it may even end up in the dictionary one day, thereby becoming normative.

We are now in a good place to bring together the different attempts at defining reference from above:

Let there be two individuals, Ann and Bob, partaking in a conversation. Both Ann and Bob begin the conversation with their own set of real patterns. These patterns have been iteratively created through two distinct reduced dialect tables, one for Ann and one for Bob. They are reduced in the sense that these

[13] Unlike my uncle, my father, who exhibited the same peculiar 'marmot' usage, was still corrigible and ultimately saw the error of his ways. This shows that successful correction cannot necessarily be expected and involves what Lassiter (2008) calls deference. The difference between the two brothers' propensity to defer during an attempt at correction involves a complicated mixture of external, social and contextual factors (such as societal views on the relationship between age and knowledge), but also internal, psychological factors (such as propensity for stubbornness).

tables contain only the entries that they have experienced. Before I sketch the final versions of reference using these reduced tables, let us first briefly relax the assumption of successful communication.

3.6. Relaxed notion of successful communication

On an intuitive definition of successful communication, meanings need to be shared between people for a communicative act to count as successful. This can be coined in terms of speakers and hearers each identifying a particular word with a particular set of objects. Then, on a naive conception of this idea, shared meanings imply a perfect one-to-one agreement between the set of objects of the speaker and the set of objects of the hearer. But if successful communication is defined as this perfect one-to-one agreement, then the semantic heterogeneity that exists in our society implies that mere mortals cannot successfully communicate and are unable to reach a bar set that high.

So instead, we water down this idea. Rather than pretend that we can only successfully communicate with one another if our words refer to exactly identical sets of objects or ideas, I would like to suggest that they only have to be similar enough so that any inevitable misalignment is irrelevant. For example, when Oscar would ask Twin Oscar for a glass of “water”, and receives a glass of XYZ, neither him nor his twin would think they miscommunicated. In contrast, my uncle’s “marmot” deviated too far from my use of the word, at least initially, and thus did not escape detection. This definition of successful communication is subjective - or at least relative to arbitrary similarity standards - because it crucially depends on what counts as sufficiently similar in the conceptual framework of the judge of the communicative act. Take Twin Earth: thanks to

Putnam's revelations, we are all-knowing observers of Oscar and Twin Oscar, capable of noticing the difference between H_2O and XYZ . For many philosophers who buy the argument, it is easy to see that they are talking about different waters - even if the Oscars are not aware of it themselves. Other philosophers argue that H_2O and XYZ are in most respects so similar (to the Oscars) that the microstructural differences should be treated as irrelevant (Crane, 1991; Segal, 2000).

Both types of philosophers have a point, and the fact that the debate about Twin Earth has raged for so long, despite there being no disagreements about any facts of the matter, should tell us that this debate is impossible to settle objectively. Whether Oscar and Twin Oscar can successfully communicate about water depends on the types of properties one finds important, and on which explanatory goal one has. It depends on arbitrary similarity measures, in this case whether the behaviours of Oscar and Twin Oscar are similar enough, and on whether H_2O and XYZ are similar enough. Like reference then, whether someone successfully communicates is dependent on an arbitrary choice between multiple, equivalently predictive, real patterns. It is therefore more useful to model this in a pluralist way: a communicative act between a speaker Ann and a hearer Bob is successful according to some judge J (where J can be Ann, Bob, or another overhearer) if and only if the meanings held by Ann and the meanings held by Bob are sufficiently similar so that any differences are judged to be irrelevant by J.

Why should there be a judge, one may ask. Should it not be possible to define the similarity of meanings between Ann and Bob objectively rather than

subjectively? By introducing a judge, am I not inviting the spectre of Fregeanism? Yes, and yes. There are two reasons for this move. First, to objectively decide whether the meanings of Ann and Bob sufficiently coincide, we are required to make a set of semi-arbitrary decisions about object-similarity, sound-similarity, etc. In other words, it requires the very same real patterns used to determine reference in the first place, and thus, like reference, there will be many ways to determine similarity of meanings. One could in principle objectively define similarity of meanings relative to a particular set of real patterns, like we did in section 3.4. But from this formulation, it is but a small step to the subjective real patterns found in the reduced meaning tables that we actually possess (section 3.5). Secondly, the move allows us to model actual conversations and semantic interactions between two subjects, all the while making room for their semantic autonomy, heterogeneity and their propensity for semantic change. We shall discuss this now.

3.7. Reference in practice

We can now finally formulate how two individuals can successfully communicate and refer to objects from the point of view of lossy semantics. We shall start with Plan A:

Let us imagine that at time t_1, Ann makes a request to Bob by uttering U. This request is for an object corresponding to Ann's pattern P_A. Bob then matches utterance U to a pattern in his memory, P_B, and gives an object O that matches his pattern to Ann. Ann compares object O to her pattern P_A. Now there are two cases: Ann can judge O to be similar to P_A, or she can judge that O is dissimilar to P_A. In the first case, following our definition of successful communication, we can

say that plan A has succeeded and that Ann and Bob have had a successful communicative interaction (according to Ann and Bob).[14]

Alternatively, Ann can judge O to be dissimilar to P_A, and choose to not accept Bob's offer of O. This initiates plan B. Ann and Bob would then initiate a negotiation about what U means.[15] This starts with an acknowledgement that their patterns P_A and P_B are different and proceeds to an attempt at creating ad-hoc patterns explaining each other's semantic behaviour. Let us assume that Ann wins this negotiation and that Bob changes his pattern P_B to be closer to P_A, at the very least excluding O from the reference of U.[16] Bob at t_2 will then change his behaviour and give a new object based on his new pattern. If Ann judges this new object to correspond to her pattern P_A, she will believe that Ann and Bob had a successful communicative interaction.

Note that in lossy semantics, the reference of this utterance U is pluralist, and time-relative. Pluralist because Ann initially disagreed with Bob about whether U referred to O. Time-relative because Bob at t_2 has a different opinion on whether U referred to O than Bob at t_1 does.

Note further that lossy semantics is nothing more than a common sense notion of how people think communication works. It is almost too trivial to write down. This theory of meaning does not expect us to know any uniquely

---

[14] There could of course be a third person, Cas, who - based on patterns P_C - judges the conversation as unsuccessful, even if Ann and Bob believe it succeeded. This is similar to a Twin Earth case where a subset of philosophers of language would reasonably judge a "water"-conversation between Oscar and Twin Oscar as an unsuccessful attempt at communication, whereas the two Oscars themselves would never be able to perceive any differences between their "water"-utterances.

[15] This negotiation can be explicit, but often is not (Pickering & Garrod, 2004)

[16] Bob could either change his pattern only when he is talking to Ann - similar to how I only changed my meaning of "marmot" when talking to my uncle. Or, he could update his general pattern, thereby changing his meaning of the utterance for all speakers. This is an example of how the way we group people into communities matters for how we update our patterns.

identifying descriptions, nor do we need to trace back the causal-historical chains to know the meaning of our words, or to find what we refer to in elaborate 2D matrices filled with a multitude of (im)possible worlds. All we require is the ability to create a set of good-enough real patterns and the flexibility to adjust these when the immediate circumstances demand it. I have no doubt that we fulfil these requirements - which cannot be said of the requirements of previous theories of meaning. Trivial though it may be, there are profound consequences for the role notions like meaning and reference play in philosophy of language. The framework shows that meaning and reference themselves do not truly exist, except as useful abstractions or real patterns of our linguistic behaviour. This makes these notions less substantial and fundamental as philosophers of language may have long thought.

What could be said however is that lossy semantics is incomplete as a theory of reference. The theory does not specify what exactly the real patterns are, nor how they should be made, and so does not tell us how we should identify a referent given a particular utterance. This is deliberate. Any theory of reference capable of capturing our behaviour is more likely to succeed if based on the study of said behaviour. In other words, determining reference stops being the job of philosophy and starts being the job of (psycho)linguistics. After all, by studying larger parts of our idiolect and dialect tables through empirical efforts, (psycho)linguists should be able to extract realer real patterns. When (psycho)linguists have figured out what types of real patterns are used by people to understand each other's references, and how exactly they learn to extract patterns from data, only then will we have a complete theory of reference.

The adjective “lossy” in lossy semantics refers to lossy compression, which implies a loss of information when abstracting a coarse-grained description from a more detailed, fine-grained description. But the framework actually involves lossy compression in two distinct ways: First, a pattern does not have to carve nature at its joints. Should natural kinds exist, then the patterns involved in the use of natural kind terms need only be good enough to reliably match roughly similar objects, which may or may not correspond to real natural kinds. The same goes for patterns among sounds, bodily movements and people. There is therefore typically some loss of information if we compare an individual’s real pattern to the actual way the world is out there. The second way the framework involves lossy compression is in relation to communication between people. The patterns of two individuals need not be exactly identical. They need only be similar enough to avoid detection. There is thus some loss of information between people when they attempt to communicate, that is, communication is lossy. The two ways of compression water down the importance of successful reference to the world and successful communication respectively.

To some philosophers this might feel like we are giving up on a way to determine reference entirely. They might worry that this means that determining reference is impossible. That there is no fact of the matter when it comes to reference (Quine, 2013). To some extent they would be right. Except perhaps in certain demonstrative situations, it is indeed impossible to identify a clear and unambiguous set of objects as *the* extension of a given utterance. We communicate using a multitude of different lossy real patterns instead.

But this is not to give up all objectivity. Despite the many difficulties with determining reference, if we are willing to play god, it remains possible to say objectively true things about meaning: the complete idiolect and dialect tables[17] contain all of the information we need when it comes to determining how (for any given person at a time t) a given utterance maps onto a given object. We can even try to come up with all kinds of patterns. A veritable rainforest filled with real patterns that are based on a great multitude of different criteria, whether they be descriptions, causal-historical chains, 2D-matrices or something else entirely that might optimally compress and roughly capture what a word should mean (Ross, 2004). We can also imagine that (psycho)linguists may attempt to measure and analyse parts of these idiolect and dialect tables more thoroughly. All of this remains possible. As long as we do not forget that any single such pattern - any one lossy compression that we may find in these tables - remains a somewhat arbitrary simplification of an incredibly complex system. And that these "optimal" simplifications will probably be rather different from the much simpler and more unstable patterns that mere mortals will actually be using to communicate with one another.

4. Conclusion

We are now in a better position to answer the four questions related to the story of my uncle's "marmots". First, how can my uncle's use of the word "marmot" be different from most people's "marmot" usage? The answer is simple: he created a

[17] These are not entirely objective, since producing the tables requires making semi-arbitrary assumptions. So more accurately would be to say that the physical bitmap on which these tables are based contains the actual objective information about meaning.

different perceived pattern than most people. As a result, his idiolect table looks rather different from the average person's idiolect table.

How was I then nevertheless able to understand my uncle's "marmot" utterance? Easy, after noticing that my uncle's use of "marmot" deviated, I stopped using my own - more established - meaning and tried to form a new pattern based on my uncle's idiosyncratic behaviour. The pattern that my uncle's "marmot" meant the animals I would call "guinea pigs" was most in line with the available evidence.

How could the truth values and inferences related to my uncle's "marmot"-utterances be affected, that is, how could "marmot" genuinely seem to refer to guinea pigs? The primary importance of comprehending reference is identifying the objects and persons our interlocutors are talking about, so we can evaluate whether what they are saying is true and are able to interact with the right objects. By rebranding reference as dynamically shifting heterogeneous real patterns in our linguistic behaviour - rather than codified one-size-fits-all word-object relations - we can say that my uncle genuinely refers to guinea pigs and is saying true things about them. Despite him using the "wrong" words, we can call this genuine reference because his consistent behaviour indicates a word-object pattern that is real, thus allowing us to identify what he is talking about.

Finally, how did my understanding of "marmot" rapidly change in order to understand my uncle? What happened was that I detected a difference between my uncle's use of the word "marmot" and my own, and was able to radically shift my linguistic behaviour towards associating "marmot" with guinea pigs rather

than marmots. While the exact mechanism is not understood, it should be clear that we possess some mechanism capable of very quickly noticing differences in word meanings (Pickering & Garrod, 2004). Moreover, we have the means to flexibly change our word meanings and related inferences. This change in us is real as it reflects a change in our own idiolect table from one moment to the next.

In summary, a good theory of meaning needs a way to deal with semantic autonomy and heterogeneity. Existing theories cannot do so properly because their strict desires to explain actual communication and true reference to the world cannot be jointly satisfied using only a few degrees of freedom. By relaxing the assumption that we refer to the world and the assumption that meaning is shared, we can explain semantic heterogeneity, while still allowing ourselves to talk loosely of shared reference to the world. The resulting theory of lossy semantics is based on the individual's ability to extract Dennettian real patterns from small samples of perceived behaviour. This ability, coupled with the semantic autonomy to radically and swiftly change our meanings, helps us to understand other people. In its foundations, lossy semantics is significantly more individualistic and dynamic than existing theories of meaning, all while retaining the ability to explain how we talk about the world and to each other.

5. Bibliography

Bird, A., & Tobin, E. (2022). Natural kinds. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2022). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2022/entries/natural-kinds/

Bjerring, J. C. (2013). Impossible worlds and logical omniscience: An

impossibility result. *Synthese*, *190*(13), 2505–2524. https://doi.org/10.1007/s11229-011-0038-y

Bjerring, J. C., & Schwarz, W. (2017). Granularity problems. *The Philosophical Quarterly*, *67*(266), 22–37. https://doi.org/10.1093/pq/pqw028

Bjerring, J. C., & Skipper, M. (2019). A dynamic solution to the problem of logical omniscience. *Journal of Philosophical Logic*, *48*(3), 501–521. https://doi.org/10.1007/s10992-018-9473-2

Bowern, C. (2019). *Semantic change and semantic stability: Variation is key* (arXiv:1906.05760). arXiv. http://arxiv.org/abs/1906.05760

Burge, T. (1979). Individualism and the mental. *Midwest Studies in Philosophy*, *4*, 73–121. https://doi.org/10.1111/j.1475-4975.1979.tb00374.x

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford Paperbacks.

Chalmers, D. J. (2004). Epistemic two-dimensional semantics. *Philosophical Studies*, *118*(1/2), 153–226. https://doi.org/10.1023/B:PHIL.0000019546.17135.e0

Chalmers, D. J. (2006). Two-dimensional semantics. In E. Lepore & B. Smith (Eds.), *The Oxford Handbook to the Philosophy of Language*. Oxford University Press.

Chalmers, D. J. (2011). The nature of epistemic space. In A. Egan & B. Weatherson (Eds.), *Epistemic Modality* (pp. 60–107). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199591596.003.0003

Courant, R., & Hilbert, D. (2013). *Methoden der mathematischen Physik*. Springer-Verlag.

Crane, T. (1991). All the difference in the world. *The Philosophical Quarterly*, *41*(162), 26.

Davidson, D. (1986). A nice derangement of epitaphs. *Philosophical Grounds of Rationality: Intentions, Categories, Ends*, *4*, 157–174.

Davidson, D. (1987). Knowing one's own mind: In *American Philosophical Association Centennial Series* (pp. 389–409). Philosophy Documentation Center. https://doi.org/10.5840/apapa2013158

Dennett, D. C. (1987). *The intentional stance*. MIT press.

Dennett, D. C. (1991). Real patterns. *The Journal of Philosophy*, *88*(1), 27–51.

Devitt, M., & Sterelny, K. (1999). *Language and reality: An introduction to the philosophy of language* (Issue 150). Wiley.

Donnellan, K. S. (1970). Proper names and identifying descriptions. *Synthese*, *21*(3/4), 335–358.

Dupre, J. (1981). Natural kinds and biological taxa. *The Philosophical Review*, *90*(1), 66. https://doi.org/10.2307/2184373

Evans, G., & Altham, J. E. J. (1973). The causal theory of names. *Proceedings of the Aristotelian Society, Supplementary Volumes*, *47*, 187–225.

François, A. (2014). Trees, waves and linkages: Models of language diversification. In Bowern, C. Evans, & Bethwyn (Eds.), *The Routledge Handbook of Historical Linguistics* (pp. 161–189). Routledge. https://doi.org/10.4324/9781315794013.ch6

Frege, G. (1879). Begriffsschrift, a formula language, modeled upon that of arithmetic, for pure thought. *From Frege to Gödel: A Source Book in Mathematical Logic*, *1931*, 1--82.

Hadamard, J. (1932). Le problème de Cauchy et les équations aux dérivées partielles linéaires hyperboliques. *Paris*, *11*, 243–264.

Haugen, E. (1966). Dialect, language, nation. *American Anthropologist*, *68*(4), 922–935. https://doi.org/10.1525/aa.1966.68.4.02a00040

Hintikka, J. (1975). The intentions of intentionality. In *Rudolf Carnap, Logical Empiricist* (pp. 183–216). Springer.

Hock, H. H., & Joseph, B. D. (2019). Language history, language change, and language relationship. In *Language History, Language Change, and Language Relationship*. De Gruyter Mouton.

Jackson, F. (1998). *From metaphysics to ethics: A defence of conceptual analysis*. Clarendon Press ; Oxford University Press.

Jackson, F. (2004). Why we need A-intensions. *Philosophical Studies*, *118*(1/2), 257–277. https://doi.org/10.1023/B:PHIL.0000019548.04704.69

Kaplan, D. (1989). Demonstratives. In *Themes from Kaplan* (pp. 481–563). Oxford University Press.

Kripke, S. A. (1980). Naming and necessity. In D. Davidson & G. Harman (Eds.), *Philosophical Books* (Vol. 22, Issue 1). Harvard University Press. https://doi.org/10.1111/j.1468-0149.1981.tb02709.x

Langston, K., & Peti-Stantic, A. (2014). Language and identity: Theoretical and conceptual framework. In *Language planning and national identity in Croatia*. Springer.

Lassiter, D. (2008). Semantic externalism, language variation, and sociolinguistic accommodation. *Mind & Language*, *23*(5), 607–633. https://doi.org/10.1111/j.1468-0017.2008.00355.x

Ludlow, P. (2014). *Living words: Meaning underdetermination and the dynamic lexicon* (First edition). Oxford University Press.

Oxford University Press. (1989a). Gay. In *Oxford English Dictionary* (Vol. 3).

Oxford University Press. (1989b). Guy. In *Oxford English Dictionary* (Vol. 3).

Oxford University Press. (1989c). Nice. In *Oxford English Dictionary* (Vol. 3).

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169–190. https://doi.org/10.1017/S0140525X04000056

Pickering, M. J., & Garrod, S. (2006). Alignment as the basis for successful communication. *Research on Language and Computation*. https://doi.org/10.1007/s11168-006-9004-0

Putnam, H. (1975). The meaning of 'meaning'. *Philosophical Papers*, *2*.

Quine, W. V. O. (2013). *Word and object*. MIT Press.

Ross, D. (2004). Rainforest realism: A Dennettian theory of existence. In D. Ross, A. Brooks, & D. Thompson (Eds.), *Dennett's Philosophy: A Comprehensive Assessment* (pp. 147–168). MIT Press.

Russell, B. (1905). On denoting. *Mind*, *14*(56), 479–493.

Schmidt, J. (1872). *Die Verwandtschaftsverhältnisse der indogermanischen Sprachen*. Hermann Böhlau.

Schroeter, L. (2003). *Gruesome diagonals*. Ann Arbor, MI: Michigan Publishing, University of Michigan Library.

Schroeter, L. (2021). Two-dimensional semantics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021). Metaphysics Research Lab, Stanford University.

https://plato.stanford.edu/archives/fall2021/entries/two-dimensional-semantics/

Schuchardt, H. (1885). *Über die Lautgesetze: Gegen die Junggrammatiker*. Oppenheim.

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

Segal, G. M. (2000). *A slim book about narrow content*. MIT Press.

Stalnaker, R. C. (1978). Assertion. *Pragmatics*, 315–332. https://doi.org/10.1163/9789004368873_013

Strawson, P. F. (1950). On referring. *Mind*, *59*(235), 320–344. JSTOR.

Strawson, P. F. (2002). *Individuals*. Routledge.

Wittgenstein, L. (2010). *Philosophical investigations*. John Wiley & Sons.

Appendix

Two-dimensional semantics

Two-dimensionalism is an attempt at synthesising Fregeanism and Millianism. It combines the Fregean and Millian ideas as two separate dimensions within the framework of a possible world semantics, allowing the two-dimensionalist to capture both the Fregean and the Millian intuitions. Initially, it was intended as a supplement and only used to explain word types that were left behind in both previous theories (Kaplan, 1989). Most prominently were the indexicals, like "I" and "she", which received their meaning not from uniquely identifying descriptions or causal-historical chains and the like. Instead, indexicals received their meaning only in the immediate context and thus needed the more flexible two-dimensionalist framework to deal with this. Later philosophers like Stalnaker (Stalnaker, 1978) expanded the framework's use to explain word meanings from the point of view of individuals, thereby explaining semantic heterogeneity. More recently, two-dimensionalism has been generalised to all utterances - including natural kinds (Chalmers, 2004, 2006; Jackson, 1998, 2004). This was done in order to additionally capture the difference between *a priori* and *a posteriori* necessary statements, which was difficult to do for the Millians.

To properly explain the two-dimensional framework, it is important to understand this difference. The difference can be illustrated by the sentences 'water = H_2O' and 'water = water'. If 'water' refers directly to H_2O, as the Millian wants, then the two statements should both reflect the very same necessary *a posteriori* truths. However, according to philosophers like Chalmers (1996) and Jackson (1998), the sentences are rather different: we had to empirically discover

the truth of the first statement, that water is H_2O. This is not true for the other identity relation: we do not need to find out that water is water. Consequently, the first statement is necessary *a posteriori*, while the second is necessary *a priori*, resulting in a difference in meaning which Millians cannot explain. Moreover, it seems as if there is still a sense in which it is contingent that 'water = H_2O'. It seems that way both i) because we could have named H_2O differently, and ii) because 'water' could still turn out to be something other than H_2O, say XYZ, should our scientific theories change. This different sense of 'water = H_2O' is again not explainable without senses or descriptions. Chalmers and Jackson thereby concluded that we needed two notions of necessity.

Necessity is typically captured by a possible world semantics, where a sentence is necessarily true if it is true in all possible worlds. The two notions of necessity therefore implied that we require two different ways to evaluate possible worlds. That is, we need at least two dimensions, one for each type of possible worlds. The framework can be best explained along with Table 2:

Table 2

Depiction of the Meanings of an Utterance of "water" according to Two-Dimensional Semantics

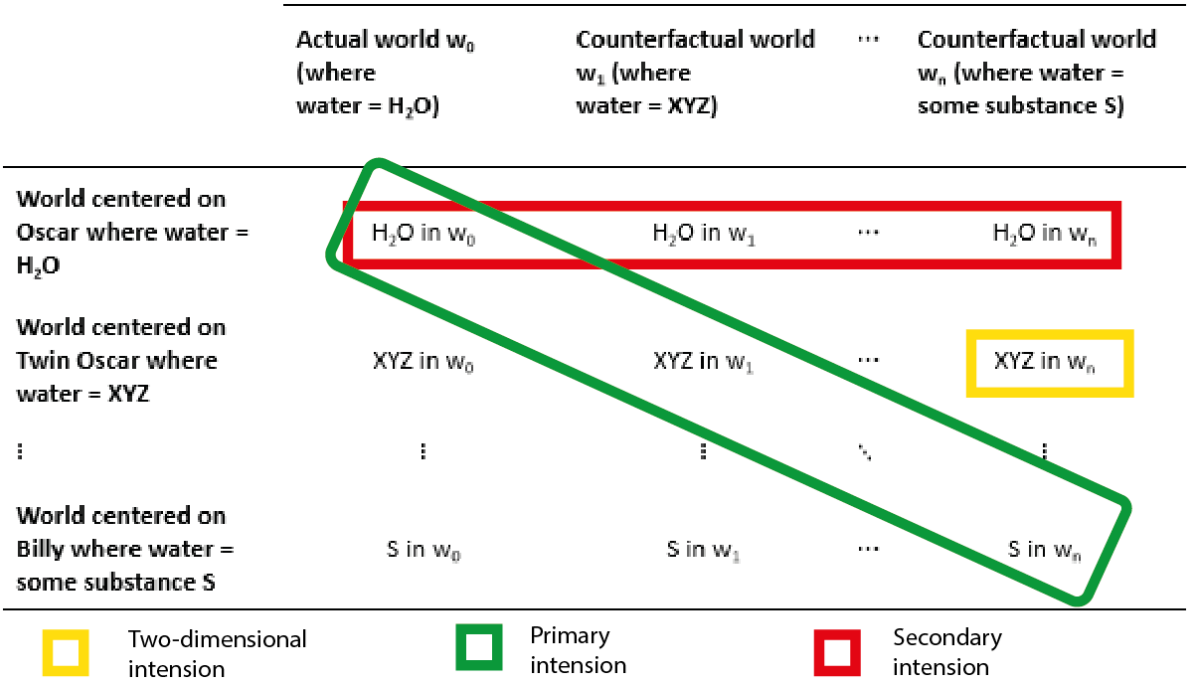


Table 2 contains two dimensions, a horizontal one and a vertical one. The vertical dimension consists of different centered worlds: different possible worlds that are centered on a particular individual at a time in that world. In Chalmers' framework, these reflect different epistemic possibilities: different ways the world could be given everything a particular individual at a particular time knew. This makes two-dimensionalism a form of semantic pluralism, allowing it to capture multiple different meanings of the same term, one per centered world. The horizontal dimension consists of different counterfactual worlds that reflect ways the world could have been. These reflect different metaphysically possible worlds.

This gives two-dimensionalism the tools to deal with different counterfactual situations that might obtain.

Chalmers then defines three intensions, each one derived from a different way of selecting the objects found within the cells of such tables, thereby denoting different sets of objects: these are called the primary, secondary and two-dimensional intensions. We start in reverse order, with the two-dimensional intensions. One way to define an intension is to say that the extension of a term is the set of objects found in but a single cell of Table 2. That is, we provide a particular centered world (a horizontal row) and a particular metaphysically possible world (a vertical column) and we select the set of objects from that cell (corresponding to the intersection of the row and column). This intension thus denotes all of the objects in the counterfactual world that correspond to the centered world's definition of the word. Chalmers calls this intension the two-dimensional intension, but it plays only a minor role.

We could also select multiple cells from this table. We could for instance select an entire row, which corresponds to what Chalmers calls the secondary or intension: given a particular centered world where water is a particular substance, H_2O for instance, the extension picked out by the secondary intension is that substance (that is, H_2O) in every metaphysically possible world. This is also called the Kripkean intension - after Kripke's rigid designation: a word denotes the same objects with respect to all possible worlds. Finally, the primary intension selects all the objects in the diagonal of the table. It looks at what substance "water" consists of according to a particular centered world, and refers to all of that substance in that same world. Its extension then is what you get when you do

this for all centered worlds simultaneously. The primary intension is supposed to express how a term's reference depends on the world one inhabits. It simultaneously captures how Oscar refers to the H_2O in his world (because he is from an H_2O Earth), while his twin refers to the XYZ in his world (because he is from an XYZ world). Because of this, the primary intension is also connected to a term's cognitive significance.

Having defined all of the terms, we can now show how two-dimensional semantics works in action. Recall that one of the main aims of generalised two-dimensionalism was to capture the difference between the necessary *a posteriori* and necessary *a priori*. This desire lead some two-dimensionalists like Chalmers (2004) to the following two claims about the relationship between intensions and necessity:

i) A sentence is metaphysically necessary if and only if its *secondary* intension is true in all (metaphysically) possible worlds.
ii) A sentence is a priori or epistemically necessary if and only if its *primary* intension is true in all (centered) worlds.

Together, these claims allow the two-dimensionalist to explain so-called necessary *a priori* sentences - a phenomenon that the two-dimensionalists felt was missing in Millian accounts of meaning. Take for instance the difference between the sentences "water is H_2O" and "water is water". These are necessary in different ways because their primary and secondary intensions are true in all worlds in different ways. "Water is H_2O" is necessary *a posteriori* because the secondary intension of the word 'water' refers to H_2O in all *counterfactual* possible worlds: it refers to H_2O along the entire row of counterfactual possible worlds, and it does so only after *a posteriori* finding out which worlds are

metaphysically possible; Instead, “water is water” is necessary *a priori* because the primary intension of “water” picks out the “water”-extensions along the diagonal, and does this for all *centered* worlds. In different centered worlds, “water” could consist of different possible substances (H_2O, XYZ, …, S). Since the primary intension of “water” always denotes exactly the substance that was used in defining “water” in that particular centered world, we know for sure and *a priori* that “water is water”. In sum, two-dimensional semantics can capture the meanings of and differences between the two “water”-statements.

Two-dimensionalism is not without its own share of problems (see Chalmers, 2006 or Schroeter, 2021 for nice overviews). Important for our purposes is that it has recently become apparent that two-dimensionalism does not completely succeed in portraying the beliefs of individuals without substantial changes to the formalism (Bjerring, 2013; Chalmers, 2011). The root of the problem is that the formalism is used for a dual purpose: on the one hand the centered worlds should reflect possible belief systems of individuals: “[T]he primary intension of an utterance or a belief is determined by the internal state of the speaker or believer” (Chalmers, 2006, p. 200).

On this first view, differences in the primary intensions of a term are designed to reflect the differences in the beliefs of the person uttering the term (Chalmers, 2011), similar to Fregean senses. On the other hand however, these centered worlds are intended to make possible the discovery of *a priori* necessary truths such as mathematical truths through conceptual analysis.

On this second conception, centered worlds sketch what is epistemically possible. Statements can be considered *a priori* necessary when they are true in all

centered worlds. Thus, a philosopher who intends to use conceptual analysis to find *a priori* necessary truths like "triangles have three corners" and "water = water", need only determine whether there are centered worlds in which the statement of interest is false. If no such worlds can be found, the statement is *a priori* necessary.

The two desires of using primary intensions i) as possible belief systems, and ii) as epistemic possibilities cannot both be fulfilled (Schroeter, 2003). It is easy to see why: people regularly have beliefs about impossible objects. Take the child who does not know of the difference between H_2SO_4 and H_2O: would its beliefs be located in centered worlds where every H_2SO_4 molecule literally *is* H_2O. Or consider someone who believes in flat Earth theory: does his centered world contain a flat Earth, even if an Earth like that would be breaking several laws of nature? What about someone who entertains ideas of square triangles? Does the possibility of entertaining these thoughts create impossible worlds where objects can somehow have both three and four corners at the same time? Finally, what should we think of the pre-1897 physicist, who still entertains the hypothesis that electrons do not exist? Does his seemingly reasonable hypothesis create a possible world where there are no electrons, even if contemporary physics holds that the macroscopic world would look so radically different without electrons, such that even the molecules of the physicist's body would rapidly cease to hold together?

In each of these cases, the two-dimensionalist is tempted to deny the existence of such a possible world, given its contradictory nature and the two-dimensionalist desire for capturing an *a priori* necessary statement as the set

of all possible worlds. Yet, the beliefs and utterances of these people reflect real informational content which would be left unexplained without the existence of centered worlds just described.

To conclude, the two-dimensionalists' insistence on modelling *a priori* necessity undercuts the viability of two-dimensionalism as a theory of reference, since people regularly say and think impossible things. Primary intensions viewed as epistemic possibilities are simply insufficiently fine-grained to capture these impossibilities. Therefore, the desire to use two-dimensionalism as an instrument to find *a priori* necessary truths is in direct conflict with the desire to use it as a behaviourally realistic theory of reference.

Bjerring (followed by Chalmers) has suggested solving the above conflict simply by adding a third dimension onto the formalism, one that indeed uses impossible worlds for its tertiary intension (Bjerring, 2013; Chalmers, 2011).[18] With these impossible worlds, Bjerring and Chalmers save three-dimensionalism as a theory of meaning, making it capable of having the fine grain required by everyday talk and beliefs, while retaining the utility of primary intensions for describing epistemic possibilities. Yet, this plan comes with an obvious cost: by duct taping this impossible third dimension onto its formalism, this new three-dimensionalism now entails an extravagant, Meinongian metaphysics not every philosopher shall be willing to accept.[19]

---

[18] Instead of three-dimensionalism, Chalmers (2011) has proposed to distinguish between an ideal and a non-ideal version of two-dimensionalism, where the ideal version's centered worlds cover the space of epistemic possibilities, and the non-ideal centered worlds cover the epistemic space plus the space of impossible worlds. Chalmer's distinction, while resulting in a slightly different formalism, essentially boils down to the same thing as three-dimensionalism.

[19] To make impossible possible worlds slightly more acceptable, Hintikka (1975) and others argue that these worlds "look possible and hence must be admissible as epistemic alternatives but which none the less are not logically possible"(Bjerring & Schwarz, 2017; Bjerring & Skipper, 2019; Hintikka, 1975). However, depending on how we read this, either this enables people to refer to the square circles that exist in these worlds, in which case there still exist impossible worlds. Or it

Two-dimensionalism fixes many of the important problems that were plaguing both Fregeanism and Millianism. I wholeheartedly agree with some of the foundational desires and principles behind two-dimensionalism - its semantic pluralism for instance. I also admire the technical, nearly mathematical, tour de force of the two-dimensionalist framework. Yet, I cannot help but view this particular product of conceptual engineering as a theoretical Frankenstein. The glueing together of the thesis of Fregeanism and the antithesis of Millianism, has not lead to genuine synthesis: the resulting two-dimensional matrix might have the appearance of unity, but by combining the Fregean and Millian ideas in the way they did, two-dimensionalists not only combined the best of both worlds, they also combined the worst. The theory is neither elegant nor parsimonious, especially in its three-dimensionalist form. While having an ugly formalism does not necessarily imply falsity, it does make a good reason for trying to find a better theory of reference elsewhere.

---

does not allow people to refer to square circles and the like, in which case the theory of reference is incomplete.

Research Proposal

1. Title of the project

Lossy Semantics

2. Summary of the theme and aim of the project (200 words)

An important outstanding scientific and philosophical question is how people use words and sentences to understand one another. To explain such mutual understanding, existing theories of meaning in philosophy of language typically assume that meanings are stable and shared. However, word meanings can change over time and people can possess different meanings, and yet they can quickly overcome their differences to understand one another. To explain these phenomena, I recently developed a new theory of meaning, called lossy semantics. Lossy semantics offers a more flexible and pluralist framework based on an understanding of meaning as Dennettian real patterns in our linguistic behaviour. Through the quick recognition of these patterns in another's behaviour, we can swiftly understand the meanings of their idiosyncratic language use.

However, lossy semantics needs to be developed further to match with the capacities of older theories to explain phenomena like compositionality and modality. The proposed research project aims to use the existing literature to incrementally synthesise a comprehensive theory of meaning. If successful, lossy semantics would be of use to more than just philosophers of language: psychologists, neuroscientists, AI researchers, linguists and society as a whole all stand to benefit from a more flexible and pluralist understanding of meaning.

3. Description of the proposed research (max 2500 words)

Background

We use words and sentences to communicate meaningfully with each other about the world. Our understanding of what meaning is remains incomplete and faulty however. The lack of a good theory of meaning is holding us back from a proper understanding of human communication in psychology and neuroscience, and is causing delays in the creation of artificial intelligence with which we can meaningfully communicate. A big gap in our current understanding is that existing theories of meaning are often idealised theories that assume that word meaning is largely homogeneous, both across time and across the individuals in a language community. Philosophers made this assumption to explain how we can communicate: a hearer can decode the arbitrary symbol into the meaning the speaker intended, because the homogenous conventional meanings of the language are shared between speaker and hearer (Lewis, 1983). In reality however, both semantic change and semantic heterogeneity are prevalent in everyday communicative interactions (Davidson, 1986; Ludlow, 2014). As a result, current theories of meaning are not equipped for explaining how word meanings change over time, and how people that use different word meanings interact and understand one another. These deficits point out that current theories have missed important aspects of what it is that allows humans to understand one another and thus do not completely capture the essence of what meaning is.

A new theory of meaning, called lossy semantics, tries to overcome these shortcomings. Contrary to previous theories of meaning, it starts from the

assumption that people can differ in what they mean by a word, but have the capacity to swiftly overcome these differences in meaning. To explain this capacity, lossy semantics is built on the idea that people can quickly extract Dennettian real patterns (Dennett, 1991) in each other's complex linguistic behaviour. And that these real patterns are what constitutes meaning. The identification of meaning with real patterns has certain implications, which alters our view of meaning.

To see why, we need to explain what real patterns are. Real patterns are Dennett's attempt to clarify the ontological status of mental states such as beliefs and desires (Dennett, 1991). According to Dennett, mental states are not strictly reducible to the physical ground truth he compares to bitmaps of a digitalized picture. Instead, mental states are like patterns in that bitmap that usefully describe our behaviour without having to go down to the low level details of that physical bitmap. These patterns are real because they contain useful information that allows us to make reasonable predictions. But they are not real in the sense that they imperfectly describe physical reality. Instead, they are lossy compressions, similar to how our computers can imperfectly reduce and lose some of the information in an image when resizing them, while still capturing the gist. The loss of less important, detailed information is what allows these patterns to generalise and describe other situations that are similar in exactly those ways captured by the pattern. An important corollary of the real pattern analogy is that there are multiple similar patterns in our behaviour that can be equally useful, and thus different observers can extract similarly useful, but different real patterns. Dennett's notion of real patterns is a useful ontological description that can be

expanded to include patterns other than mental states (Ladyman et al., 2007; Petersen, 2019; Ross, 2004). Importantly, it can be extended to include the meanings of words and sentences as lossy semantics argues. Thus, the implications of being a real pattern also apply to meaning.

The possibility of creating multiple equally useful patterns in the same bitmap is especially relevant. Since individuals often have different ideas on what a particular word means, we cannot always depend on static dictionaries that associate a word with a single meaning. While we do regularly assume that we have a shared meaning, sometimes this assumption does not hold. When this happens, we discard our usual go-to meaning. Instead, we dynamically extract real patterns in the immediate linguistic behaviour of our interlocutors that help us determine what they mean by a word, and lets us once more understand the other person. By framing meaning as a real pattern of our linguistic behaviour, we can move beyond static and homogeneous word-object mappings, and create a more dynamic, realistic theory of meaning.

Aims

However, lossy semantics is still in its infancy and requires further development to be on par with existing theories of meaning in other domains. Specifically, the current research proposal has four aims to develop the framework further. Two of these four aims correspond to important aspects of meaning that can be explained in existing theories of meaning, but which are not fully developed in lossy semantics yet: compositionality (aim 2) and modal semantics (aim 4). Developing these two aims matures lossy semantics, thereby making it a candidate for an improved replacement of existing theories of meaning. The other two aims

concern two long standing debates in philosophy of mind and language, that could potentially be resolved through lossy semantics' notion of real patterns: the internalism-externalism debate (aim 1) and the denotationalism-inferentialism debate (aim 3). Resolution of these debates would help move the field forward, and at the same time result in a more comprehensive theory of meaning.

1. Reference to the mind-independent, physical world through real patterns in partial data

Philosophers of mind and language have for a long time debated whether meaning is something internal to us (Chomsky, 2000; Mcgilvray, 1998; Pietroski, 2018) or whether it depends on something external, either the social domain (Burge, 1979) or the world itself (Putnam, 1975). The referential real patterns of lossy semantics are part internalist, part externalist. On the one hand, individuals might extract different patterns, on the other hand real patterns are real and thus dependent on external properties of the world.

One aim of the current project is to consider whether the internalist and externalist views on meaning could be seen as two sides of the same coin. The working hypothesis underlying this idea would be to view reference to the external world as restricted, but still loosely possible through the internal creation of real patterns in our linguistic behaviour aimed at the physical world (=external). Importantly, these patterns would still be created through partial data (=internal) received from the external world. By fleshing out and evaluating this hypothesis I hope to resolve a long-standing, ill-posed philosophical debate and lay more solid foundations for lossy semantics.

2. Combining real patterns into compositional meaning

Language is compositional: words combine into larger phrases and sentences (Frege, 1980). For instance, “the current king of France” has a definite meaning that is more than merely the sum of the parts. If, as lossy semantics posits, word meanings truly are referential real patterns, then if it is to explain compositionality, real patterns should be combinable. That is, referential real patterns should be able to be composed together to form phrase and sentence meanings. But what exactly does it mean to combine two patterns together? And how does the combination of multiple patterns, - for instance those related to the words “current”, “king” and “France” - when put together form a definite meaning that is distinct from the patterns of the words separately?

The second aim of the proposed research project is to develop a lossy semantics account of such compositional meaning. The hypothesis here is that the result of combining two or more patterns is itself a higher order, more abstract pattern. This can be illustrated by the example of a (digital) image of a square. Such a square is a pattern composed of four lines. Each of these lines are themselves formed through a particular combination of pixels in the bitmap. Moreover, similar to how words combine into sentences, syntax cannot be ignored, that is: it is important how exactly the lines and pixels are combined and related to each other (e.g., placing four lines in parallel will not result in a square). The importance of syntax could explain how compositional meaning is more than just the sum of its parts. While a reasonable hypothesis, the current proposal aims to test it by carefully verifying how it applies to the different varieties of compositionality found in the literature.

3. Real patterns as bridge between denotationalism and inferentialism

Contemporary philosophy of language is roughly split into two camps. In the first camp are the denotationalists, who believe that words denote things in the world (Frege, 1879; Russell, 1903). They believe that meaning is about correspondence to the world: that "snow is white" is true if snow is white, i.e. that the thing that denotes "snow" has the property that denotes "white". In the opposite camp, there are the inferentalists who argue that the representational character of language is overblown. Instead, language meaning is about what inferences are afforded by how the words and sentences relate to one another and how they are used in their social context (Brandom, 2000; Sellars & Rorty, 1970; Wittgenstein, 2010). Both theories seem to get at real aspects of language meaning, but the frameworks are so different that they have not been unified yet.

Real patterns might be the missing link between the two camps. On the one hand, the referential real patterns are still supposed to loosely denote things in the world. On the other hand, reference is a real pattern of the relationships that exist in our perceived linguistic behaviour and they are therefore eminently inferential in nature. Because lossy semantics straddles the line between these two camps, the third aim of the proposed research project is to explore whether referential real patterns might be able to synthesise denotationalism and inferentialism into a unified semantic theory.

4. Modality, possible worlds and non-existence

People can speak and think about things and situations that are not (yet) actually true, e.g. "this research proposal is approved". They sometimes even speak about

things and situations that cannot possibly be true , e.g. “the square circle is blue”. Arguably, when they do so, their words and sentences still have meaning. In fact, we use codified words such as “could”, “might”, “cannot” etc. to explicitly mark the modality of certain situations we wish to speak of. A complete theory of meaning therefore requires an explanation of such modal meanings (Kripke, 1980; Lewis, 2013).

Since lossy semantics aims to be complete, it needs to explain modality. The fourth aim of this proposal is to establish an account of modality. This might be done through the compositionality of real patterns, developed in the second part. The gist of the idea is that people can combine multiple real patterns into a composite structure that may or may not exist, e.g. they can combine the patterns associated with “circle”, “square” and “blue” into “a circle that is both square and blue”. This structure, though it does not refer to the world by itself, remains tethered to the world through its real components. In this way the composite structure retains some meaning. The goal is for lossy semantics to explain the meaning of modal expressions with this compositional mechanism, thereby extending its explanatory scope.

Methods

The proposed methodology employed is one of iterative synthesis of the relevant literature. This concerns among others the literatures of the internalism/externalism debate, the literature on compositionality, those of denotational and inferential semantics, and finally that of modal semantics. In each of these cases, the goal is to identify for the relevant theories what their explananda are: what are the phenomena that this theory - and thus a good theory

of meaning - should be able to explain? And importantly: which of these phenomena are currently not explained by the nascent lossy semantics?

The aim is then to adapt and build upon the existing more dynamic and heterogeneous framework of lossy semantics to additionally explain these phenomena. By going over one set of theories at a time, and by incrementally adjusting and expanding the theory of lossy semantics, the resulting synthesis should explain ever larger sets of explananda. The end product of this methodology should be a more consistent and comprehensive theory of meaning, capable of explaining both old and new phenomena.

Scientific and social relevance of the research project

Scientifically, there is an urgent need for a better conceptualization of meaning.

This is true for neuroscientists and psychologists, who aim to study meaning as an emergent process of our brains and our psychology, but who often employ overly static and homogeneous notions of meaning in their studies on neural representations. Because of this conception of meaning, they often make flawed assumptions, for instance that it is unproblematic to average over participants, or that meanings remain stable over the course of an experiment; it is equally true for the field of artificial intelligence, which tries to create artificial agents capable of natural language processing (Devlin et al., 2019; Radford et al., 2019).

Contemporary artificial models are great at pattern recognition but are often trained on goals only tangentially related to meaning, namely to predict the next word in a sentence. A successful theory of meaning could redefine these learning goals and propose which semantic patterns are worth detecting, potentially leading to better models. Finally, it is true for both philosophers of language and a

large subset of linguists, who still maintain largely idealised, homogeneous pictures of language that could likewise benefit from more realistic conceptualizations of meaning.

Besides the trickle down effects of better scientific methodologies and models - especially of better natural language processing in artificial intelligence, there are other implications. The main societal relevance of a clearer understanding of dynamic, heterogeneous meaning lies in the development of more effective ways to prevent and resolve inevitable miscommunication - which is especially prevalent in an age of large scale inter-bubble communication. A simple improvement would be to inform society that people might differ in what they mean by a term, what the effective ways to detect semantic differences are and how to overcome them. For instance, if the working hypothesis developed as part of aim 2 bears fruit, then more abstract patterns are composed of less abstract ones. If so, then a successful strategy of detecting and overcoming semantic differences could be to decompose terms into their simpler constituents. This is already what we implicitly do when we provide definitions for our terms or when we point out particular exemplar objects, but lossy semantics could generalise this idea and offer a framework for identifying what constitutes “simpler” here. Moreover, its connection to patterns might leverage insights in the pattern recognition, statistics and information theory literatures to gain more efficient ways to explore the space of possible patterns. At the very least, lossy semantics’ emphasis on plural and dynamic meanings could inform the interested layman in what the mechanisms behind miscommunication are and how static and singular meanings are rarer than we often think.

4. Keywords

Meaning, Communication, Real Patterns, Semantic Pluralism, Semantic Change, Philosophy of Mind and Language

5. Timetable

The plan is for every aim to result in a separate published paper, preferably published in *Mind and Language*.

After 12 months:

Achieve the first aim: Reference to the mind-independent, physical world through real patterns in (sense) data

After 24 months:

Achieve the second aim: Combining real patterns into compositional meaning

After 33 months:

Achieve the third aim: Real patterns as bridge between denotationalism and inferentialism

After 42 months:

Achieve the fourth aim: Modality, possible worlds and non-existence

After 48 months:

Fully written PhD thesis, with introduction, discussion and summary.

6. Summary for non specialists (500 words)

To communicate with one another, humans use language: arbitrary symbols and sounds that have meaning, often referring to things in the world. Contemporary theories of meaning in philosophy of language often assume that words are used homogeneously within a particular language community, both across time and across people. These assumptions restrict current theories: they cannot properly

explain how words change meaning over time and how people that do not share the same word meanings are still able to communicate. To improve the current state of the art, a new theory called lossy semantics does away with a notion of reference defined as a static, singular word-object relationship. In its place, it proposes that meaning is a real pattern in our linguistic behaviour. That is, a language community often uses a given word to refer to similar types of objects.

Unfortunately, no one is always entirely consistent. Sometimes individuals have learned the wrong word-object relationship, or fail to identify the object correctly, and thus behave differently from the established norms. Other times, the meaning of the word is in flux, and a subset of the language community starts to use a word differently from the rest. Thus, while there generally is some pattern to our behaviour, it is not a perfect pattern and sometimes fails to explain and predict what people mean by a given word. In these cases, we need to dynamically read the other's immediate behaviour to detect new patterns that tell us what their idiosyncratic meanings are.

Lossy semantics offers a way to think about meaning in a more flexible and diverse manner, and therefore is a step up from older theories of meaning. However, since it is still new, it does not have all of the old features more established theories have. It cannot yet account for the compositionality of language - how words combine into larger phrases and sentences - nor can it explain the meanings of modal expressions - those involving terms like "could", "must", etc. By building upon and combining the insights of older theories of meaning - even those that were previously at odds with one another - lossy semantics might be able to become a unified theory of meaning. If this bears out,

then this would make improvements to our conceptualization of meaning that could be of use to other scientific disciplines such as psychology, neuroscience, artificial intelligence and linguistics. In addition, a more flexible and diverse understanding of language meaning might allow society to more quickly resolve debates and misunderstandings that revolve around semantic differences.

7. Bibliography

Brandom, R. (2000). Articulating reasons. In *Articulating Reasons*. Harvard University Press.

Burge, T. (1979). Individualism and the mental. *Midwest Studies in Philosophy*, *4*, 73–121. https://doi.org/10.1111/j.1475-4975.1979.tb00374.x

Chomsky, N. (2000). *New horizons in the study of language and mind*. Cambridge University Press.

Davidson, D. (1986). A nice derangement of epitaphs. *Philosophical Grounds of Rationality: Intentions, Categories, Ends*, *4*, 157–174.

Dennett, D. C. (1991). Real patterns. *The Journal of Philosophy*, *88*(1), 27–51.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

Frege, G. (1879). Begriffsschrift, a formula language, modeled upon that of arithmetic, for pure thought. *From Frege to Gödel: A Source Book in Mathematical Logic*, *1931*, 1--82.

Frege, G. (1980). *Philosophical and mathematical correspondence*. Blackwell.

Kripke, S. A. (1980). Naming and necessity. In D. Davidson & G. Harman (Eds.), *Philosophical Books* (Vol. 22, Issue 1). Harvard University Press.

https://doi.org/10.1111/j.1468-0149.1981.tb02709.x

Ladyman, J., Ross, D., Collier, J., Spurrett, D., Spurrett, D., Collier, J. G., & others. (2007). *Every thing must go: Metaphysics naturalized*. Oxford University Press on Demand.

Lewis, D. (1983). *Philosophical Papers Volume I*. Oxford University Press. https://doi.org/10.1093/0195032047.001.0001

Lewis, D. (2013). *Counterfactuals*. John Wiley & Sons.

Ludlow, P. (2014). *Living words: Meaning underdetermination and the dynamic lexicon* (First edition). Oxford University Press.

Mcgilvray, J. (1998). Meanings Are Syntactically Individuated and Found in the Head. *Mind & Language*, *13*(2), 225–280. https://doi.org/10.1111/1468-0017.00076

Petersen, S. (2019). Composition as pattern. *Philosophical Studies*, *176*(5), 1119–1139.

Pietroski, P. M. (2018). *Conjoining meanings: Semantics without truth values*. Oxford University Press.

Putnam, H. (1975). The meaning of 'meaning'. *Philosophical Papers*, *2*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., & others. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Ross, D. (2004). Rainforest realism: A Dennettian theory of existence. In D. Ross, A. Brooks, & D. Thompson (Eds.), *Dennett's Philosophy: A Comprehensive Assessment* (pp. 147–168). MIT Press.

Russell, B. (1903). *The principles of mathematics*. Cambridge University Press.

Cambridge-UK.

Sellars, W., & Rorty, R. (1970). Science and metaphysics: Variations on Kantian themes. *Philosophy*, *45*(171), 66–70.

Wittgenstein, L. (2010). *Philosophical investigations*. John Wiley & Sons.

8. Curriculum Vitae PhD candidate

Personal information

Name	Rowan Sommers
Address	Kerkpad 87, 6543 XN, Nijmegen
Phone nr.	0636270514
E-mail	rowansommers@gmail.com
Nationality	Dutch

Research experience

Sep 2017 – Now — **PhD student**

Lab*:* **Neurobiology of Language**

PI*:* **Peter Hagoort**
Supervisors*:* **Mante Nieuwland, Jan-Mathijs Schoffelen**
Institute*:* **Max Planck Institute for Psycholinguistics**
Topic*:* **Neurobiology of Reference**

Language comprehension requires making a mapping from words/phrases to objects in our multimodal sensory world. In my PhD I investigated using EEG, MEG and computational modelling how the brain is able to make this mapping, with a focus on how it is able to learn it as fast as it does, i.e., in one-shot. Since ANNs struggle to explain this, I co-developed a symbolic-connectionistic hybrid model capable of simple one-shot reference comprehension based on fast Hebbian learning.

Sep 2016 – Aug 2017 — **Master student (CNS internship)**

Lab*:* **Artificial Cognitive Systems**
PI*:* **Marcel van Gerven**
Supervisors*:* **Umut Güçlü, Sander Bosch, Katja Müller**
Institute*:* **Donders Institute for Brain, Cognition and Behaviour**

Topic: **Encoding/decoding of visual representations in a densely sampled fMRI dataset (Doctor Who dataset) using ANNs**

Visual perception is thought to consist in parallel activation of statistically learned hierarchical representations. To study the structure of these hierarchical representations, I set up, piloted and collected a large-scale multi-session single participant fMRI dataset while the participant was watching Doctor Who. I then predicted the BOLD response of voxels in the visual cortex using a convolutional neural network pretrained on ImageNet.

Other relevant experience

Nov 2020 – Oct 2021	**Member of Diversity and Inclusion working group**
Oct 2018 – Jun 2021	**Member of Social Committee NBL lab**
Nov 2018 – Aug 2019	**Co-initiator Theoretical discussion lunch meetings MPI**
Sep 2018 – Jul 2021	**PhD representative Research Facilities Committee MPI**
May 2020 – Jul 2021	**Co-initiator linear algebra + calculus study group**
Jan 2016 – Aug 2017	**Co-founder and treasurer CNS study association Dondrite**
Sep 2014 – Aug 2016	**Co-initiator Wittgenstein reading group**

Education

Sep 2013 – Now	**Research Master Philosophy** **Radboud University** Specialisation: **Philosophy of Mind and Language** Thesis: '**The meanings of meaning'** Supervisor: **Marc Slors**

Sep 2015 – Aug 2017	**Research Master Cognitive Neuroscience** (Cum Laude) **Radboud University** Specialisation: **Perception, Action and Control**
Sep 2013 – May 2014	**Interdisciplinary Honours Program Master** Topic: **Deep brain stimulation**
Sep 2011 – Aug 2014	**Bachelor Philosophy** (Bene Meritum) **Radboud University** Specialisation: **Philosophy of Mind and Language** Thesis: **'Knowledge argument: learning what it's like to see red.'** Supervisor: **Bart Geurts**
Sep 2009 – Aug 2012	**Bachelor Psychology** **Radboud University** Specialisation: **Brain** Thesis:'**Gamma-band activity modulated by anticipatory attention during sustained tactile stimulation.'** Supervisors: **Eric Maris, Freek van Ede**
Sep 2010 – Aug 2013	**Interdisciplinary Honours Program Bachelor**

Publications

Nieuwland, M. S., Coopmans, C. W., & Sommers, R. P. (2019). Distinguishing Old From New Referents During Discourse Comprehension: Evidence From ERPs and Oscillations. *Frontiers in Human Neuroscience*, *13*, 398. https://doi.org/10.3389/fnhum.2019.00398

Seeliger, K., Sommers, R. P., Güçlü, U., Bosch, S. E., & van Gerven, M. A. J. (2019). A large single-participant fMRI dataset for probing brain responses to naturalistic stimuli in space and time. *BioRxiv*. https://doi.org/10.1101/687681

Sommers, R. P., Dings, R., Neijenhuijs, K. I., Andringa, H., Arts, S., van de Bult,

D., Klockenbusch, L., Wanningen, E., de Bruin, L. C., & Haselager, P. F. G. (2015). A Young Scientists' Perspective on DBS: A Plea for an International DBS Organization. *Neuroethics*, *8*(2), 187–190.

Competences

Data acquisition	EEG, MEG, fMRI
Data analysis	Neural networks (mostly CNNs and LSTMs), GLM, Mixed effects models, Bayesian statistics, cluster-based permutation tests, EEG/MEG time-frequency analysis, MEG source reconstruction, fMRI pre-processing and decoding
Programming	Python, MATLAB, R, Presentation, Bash
Mathematics	Calculus, linear algebra, probability theory, information theory, set theory, dynamical systems theory
Languages	Dutch (native), English (fluent), German (basic)

Other interests

I have spent a large chunk of the last couple of years renovating my old house into a liveable place. My fiancé thinks I am an "OCD-er": obsessive compulsive denker (thinker). Just like how some people like to learn other natural languages, I like to learn other scientific languages, for they often give me new tools for thinking.