Manifolds as conceptual representations in formal semantics

Michael Goodale

June 7, 2022

Declaration of originality

This work aims to address the representations of concepts in formal semantics. While this is well-trod ground for many different theories of content words, this thesis takes a different approach then which is generally taken.

Specifically, it integrates different techniques from machine learning, differential geometry and cognitive psychology as part of the structure of concepts. While there have been previous attempts to wed machine learning techniques to formal semantics (Baroni, Bernardi, and Zamparelli 2014) or psychological theories to semantics (Del Pinal 2015), these have taken a totally different approach. This will be shown to produce a semantic model which can be implemented computationally and which provides accounts of generic sentences, certain reasoning problems and privative adjectives. It can also provide accounts for unaccounted expressions such as “fake Obama”.

Declaration of contribution

I declare that the entirety of the thesis was written by me in preparation for my Master’s degree. Excluding where explicitly acknowledged, this work is the product of my own work. No part of the thesis has been previously submitted to any publication nor in preparation for any other degree or qualification. My supervisor, Salvador Mascarenhas continually read different drafts of the work and suggested various edits.

The theoretical work behind this thesis was carried out over the course of the past year by me and informed by discussions and advice from my supervisor throughout the year. Any programming work carried out was solely my own.

Abstract

I lay the foundation of a formal theory of concepts and a formal language to manipulate concepts compositionally, unifying insights from linguistics, philosophy, psychology, and artificial intelligence.

The approach can be used in a formal semantic framework and defines concepts as mathematical objects that incorporates both extension-determining sets along with notions of similarity and typicality which have long been argued to be central to concepts. Under this view, concepts are tripartite: each concept consists of a manifold, a probability distribution over the points of the manifold, and a metric tensor to determine distances along the manifold. The manifold determines the extension of a concept over a vector space of all conceivable objects. The metric tensor quantifies the distance between objects along the manifold. This allows us to determine how similar objects are to one another *qua* a given concept. Finally, the probability distribution serves to define the typicality of a given object within a concept.

I apply this preliminary model to generic statements, privative modification, Frege's puzzle and certain reasoning tasks, showing its promise in providing accounts of diverse and recalcitrant puzzles in semantics. Finally, a computational implementation of the model using a neurosymbolic approach with invertible neural networks is described.

Contents

Chapter 1

Introduction

A central part of human cognition is our way of dividing up and categorising the world into discrete concepts. The universe does not serve neatly-defined labels to us on a silver platter nor are there straightforward, metaphysically defined laws which allow us to determine what is what. Rather, we enter the world and messily build up a repertoire of concepts over the span of our lives. These concepts help structure our language and our thought.

When we encounter something, we compare it to other things we have previously seen. From this, we categorise it and try to reason about what properties it might have. Indeed, much of a child's development is spent trying to understand the concepts that we use in day-to-day life.

While it is possible to babble about entirely abstract jabberwockies without any meaning, nearly all of our speech involves conveying complex representations which combine and contrast the concepts we are thinking about. We can deftly construct intricate novel ideas and scenarios and we are able to convey these ideas to others. Our concepts allows us to not simply manipulate strings of words according to a grammar, but actually express real *meanings* to one another.

1.1 What is a concept?

Something that has often lead to much confusion in cross-field communication about concepts is that each field and each discipline uses the word in slightly different ways. We need to precisely nail down what kinds of concepts we are interested in, and which we are not. While there has been constant back and forth for the past 2500 years of Western-philosophy about whether concepts exist solely in our head or not, this thesis develops an internalist, psychological, and linguistic theory. As such, the concepts in this work are the *representations* that allow us to carry out certain kinds of thought or speech.[1]

1.1.1 Technical concepts

A "concept" is a messy idea which can be applied from things as simple as "cups", to more complicated socio-political phenomena such as "antidisestablishmentarianism". It is often even used to describe the ability to perform novel cognitive tasks (e.g. counting).

Tasks such as counting or scientific theories are often given a great deal of attention in the psychological literature about concepts (Carey 2009). These are tasks which can take a great deal of time to acquire for children, and which are learnt in diverse, idiosyncratic ways.

Instead of these kinds of novel cognitive tasks, the work will focus on more basic concepts. For example, while the number one can be given a precise definition using the Zermelo-Fraenkel-Choice axioms, a common everyday concept like "lion" resists precise definition. While we may be tempted to develop a theory about lions (e.g. perhaps it's something's DNA that makes it a lion), it remains *remarkably* easy for someone to refer to lions or recognise lions without any knowledge whatsoever about the particularities of that lion, whether its DNA or something else. Indeed, we can even use "lion" to refer to a statue of a lion or even metaphorically to refer to a particularly brave and braggadocious individual, even though both of these things do not have lion DNA.

It is simple everyday concepts like "lion", "red", "fake" or "shoe" that this work aims to address, rather than the complicated novel conceptual systems which are learnt to do novel tasks such as counting or multiplication. In essence, concepts where we learn "what" something is rather than learn "how" something is done.

1.2 How are concepts relevant to formal semantics?

Concepts are integral to semantics if we are aiming to understand the meaning of sentences; it is important to also understand the individual words that make up that sentence (whether functional or lexical). However, most work in formal se-

[1] A useful distinction which has been proposed by Johnston and Leslie 2019 are φ-concepts and ψ-concepts: concepts as used by philosophy and psychology.

φ-concepts are often used to describe the meaning of a term on a metaphysical level or from a semantically-externalist perspective. ψ-concepts, in psychology, are not interested in the metaphysical meaning of a term; rather the computations that allow people to use it. Whether its the heuristics or folk-theories that people use to categorise things, ψ-concepts are *not* the precise formal definitions that philosophers are interested in.

mantics aims to abstract away the meaning of individual concepts, by representing concepts as sets of individuals.

$$[\![\text{lion}]\!] = \{l_1, l_2, \ldots\}$$

This abstraction allows semanticists to focus on functional words such as quantifiers or connectives. These functional words operate over the sets themselves, disregarding the way the set itself is generated.

Nonetheless, our conceptual representations keep rearing their ugly heads and interfering with the analytic, abstract work of semanticists. A variety of different linguistic phenomena seem to integrate our conceptual understanding of the extensions of words such as generic sentences or privative modification (P. Johnson-Laird 1982; Kamp and Partee 1995; Leslie 2007; Del Pinal 2015) These phenomena seem to suggest a deeper level of representation than just sets. While there are countless effects of conceptual representations on language, a subset of these linguistics phenomena seem tightly linked to the notions of similarity and typicality (Guerrini 2021). These notions are foundational to many different theories of concepts, notably prototype theory (Rosch 1975).

The particular phenomena I aim to address are generic sentences, privative adjectives, modes of presentation and indirect illusory inferences from disjunction. I give a brief overview of each phenomenon here and save the previous theoretical accounts for their respective chapters.

1.2.1 Generic statements

Generic statements express broad generalisations or typical features for a concept. In English, they are typically expressed using the bare plural construction, though in some contexts they can be licenced with the singular definite article or the indefinite article. (1) shows some typical examples of generic sentences.

(1) a. Birds fly.

b. Lions have manes.

c. Canadians eat maple syrup.

While it may seem tempting to suggest that they simply refer to cases where the predicate is true for the majority of subjects, (2) shows that this is not the case.

(2) a. Mosquitoes carry malaria.

b. # People are right-handed.

He is **French** doctor. He is a writer.	He is **skilled** doctor. He is a writer.	He is **fake** doctor. He is a writer.
∴ He is a **French** writer. ✓	∴ He is a **skilled** writer. ✗	∴ He is a **fake** writer. ✗ ∴ He is a doctor. ✗
(a) An *intersective* adjective	(b) A *subsective* adjective	(c) A *privative* adjective

Table 1.1: Entailment patterns for three types of adjectives.

The vast majority of mosquitoes do not carry malaria and most people are right-handed and yet only (2a) is appropriate.

Furthermore, the actual structure of the predicate in question seems relevant.

(3) a. Ducks lay eggs

b. # Ducks are female

(3a) is a perfectly acceptable sentence and (3b) is clearly false, despite only female ducks being capable of laying eggs.

1.2.2 Privative modification

Some adjectives are (or appear to be) *intersective*. The noun and the adjective each contribute a set and the resulting object is in the intersection of them. For example, a red apple is both an apple and red.

Subsective adjectives are a broad category of adjectives where the meaning of an adjective depends on the noun it modifies. This is typically understood as selecting a *subset* of the noun. For example, a skilled musician is a musician who is skilled at playing music, not the intersection of skilled things and musicians. For example, $\mathit{skilled}(x)$ is really $\mathit{skilledAt}(y)(x)$ where y is a free variable corresponding to what one is good at (Montague 1970; Morzycki 2015).

Finally, there are *privative* adjectives. These adjectives are similar syntactically to subsectives (i.e. they may also have a free variable), but carry an additional entailment. Specifically, they imply that the object in question is *not* a member of the relevant set. For example, a "fake doctor" is not a doctor at all.

There is also the very interesting class of intensional or modal adjectives such as 'alleged'. These adjectives often have entailment patterns similar to privative adjectives, but display different syntactic properties across multiple languages (Partee 2010). For example, they cannot occur in predicative position unlike other privatives:

(4) a. That is a fake gun.

b. The gun is fake.

(5) a. John is an alleged murderer.

b. # John is alleged.

They have different syntactic properties than privative adjectives like “fake” and are thus a different category. I will not address them in this work.

Some adjectives can also be incidentally privative, for example, a plastic lion will generally be understood as a model of a lion that is made of plastic, rather than a literal lion which is plastic.

Privative adjectives are of considerable interest because they cannot be easily accounted for by the typical set-extensional framework in formal semantics. It seems precisely to be those things which are *like* guns on some level (e.g. physical appearance), but not others (e.g. purpose). Similarity seems key to understand what things are *fake guns* and what things simply aren’t *guns* at all.

1.2.3 Reasoning and representativeness

Another issue where typicality and similarity seem to pop up again and again is the psychology of reasoning. Since Tversky and Kahnemen’s groundbreaking work in the 1970s, psychology and economics have had to reckon with the fact that people routinely make inferences and conclusions which are fallacious by normative standards of reasoning (Tversky and Kahneman 1974; Tversky and Kahneman 1983).

A huge number of these fallacious inferences seem to be related to the phenomenon of *representativeness*. That is, people assume something is likely or true because it is *typical*. For example, given a description of a feminist-sounding women, people assume that is more likely that she is a feminist and a bank-teller rather than that she is a bank-teller (Tversky and Kahneman 1974). This is a blatantly violation of the axioms of probability, yet people do it robustly. However, a feminist bank-teller is more representative or typical of the description than just any bank-teller, so people infer that “feminist bank-teller” is more *likely*.

Many fallacies can be partially accounted for on the basis of representativeness, from base-rate neglect to indirect illusory inferences from disjunction (Sablé-Meyer and Mascarenhas 2021).

1.3 Previous approaches to concepts

Across different fields in cognitive science, there have been countless proposals of theories of concepts, some of which I will outline here to highlight their strengths and weaknesses. Each discipline approaches concepts in a different way to solve different problems; but they each have independent, valuable insights which ought to be synthesised.

1.3.1 Philosophical and linguistic theories of concepts

A core puzzle of Western philosophy for over two thousand years has been the categories that we use in conversation to divide the world into its smaller constitutive parts. There are real metaphysical and ontological consequences depending on the view that one takes: are the categories we use an intrinsic part of the structure of reality, an intrinsic part of our own human thought independent of the outside world, or simply useful tools with which we talk about and interact with the external world?

Indeed, many linguistic theories of categories explicitly use the divisions of kinds drawn up by Aristotle (Pustejovsky 1995). One idea that has been put forward throughout history is that a category is defined by what we can say is true of it. The category, or concept of a "lion" then would be defined by a constellation of facts of the form "It is true that lions have manes; it is true that lions are carnivores" and so on. This notion of a concept connects it closely with the kinds of generic sentences discussed in Section 1.2.1, and was an enormous influence in twentieth century philosophy of language.

We also need to be able to *compose* concepts, that is to combine two independent concepts into a new concept. As Aristotle notes, concepts like "lion" cannot be true or false on their own: only sentences can be, and sentences come from the composition of words. Philosophers of logic and language from the turn of the twentieth century picked up on this concern for *compositionality*, most famously articulated by Gottlob Frege. In a nutshell, this is the idea that words have individual meanings and that sentences get their meaning fundamentally from how these individual meanings are put together. This philosophical principle proved foundational in linguistics, as the study of compositional semantics was born in the 1970s (Montague 1970).

Over the past fifty years, linguistic semantics, pragmatics, and philosophy of language have had many great successes building powerful and insightful theories of the meanings of complex sentences, and of function words like "only," "every," or "must." But *concept words* themselves remained deeply problematic. Somehow, it has been easier to make progress on what "most" means than on what "lion" means.

The most popular view by the 1960s was from Frege and Russell: concepts were simply shorthand for complex, long descriptions (Frege 1892; Russell 1905). For example, a lion could be a large feline predator living in Africa with a mane whereas "Aristotle" could be "the teacher of Alexander the Great who wrote *Meta-*

physica." But of course, these descriptions might not be true for every member of a concept. For example, "lions have manes" sounds true, yet not every lion has a mane.

Russell's descriptive theory of concepts was influential for about half a century before it was bitterly attacked by American philosophers Saul Kripke and Hillary Putnam. They built a very strong case against the descriptive theory of names and categories, pointing out in particular that descriptive claims about a concept might change, while the concept's members, and the concept itself, stay precisely the same, showing that the determiner of the concept cannot simply be a collection of facts (Putnam 1973; Kripke 1977; Kripke 1980).

A clear example of this is the concept of "fire." People in the Renaissance believed that fire was the product of a hitherto unseen element known as *phlogiston*. Under that out-dated scientific theory, fire would be described as the reaction of phlogiston and the air when substances leaked phlogiston. Later on, of course, we discovered this was not in fact how fire works. If the descriptive theory of concepts were truly correct, that would entail that what we thought was fire was in fact something else, since the description of fire is that it is the product of phlogiston. Instead, we concluded we were wrong about *the properties* of fire. The collection of things that are instances of "fire" remained the exact same. Consequently, many philosophers proposed an *extensional* view of concepts, where concepts are identified with the set of all elements in the universe that are instances of the concept at hand, irrespective of the existence of suitable descriptions that might define the *intensions* of those sets.

Linguistics largely inherited the philosophical view on concepts. Since semanticists are primarily interested in compositionality, the actual meaning of a concept is arguably of little importance. Accordingly, mainstream linguistic semantics has happily regarded the meanings of words like "lion" as signifying simply the set of all lions in the universe. This has allowed semanticists to give impressive accounts of things like quantification (Szabolcsi 2010).

The view is highly problematic. First, if a concept is the set of things that instantiate that concept, how can we consider novel concepts, that might not actually be instantiated in any way? Concretely, no one I know has ever met a "Harvard-educated carpenter." Yet studies in social sciences have shown that humans are happy to ascribe properties to such concepts. For "Harvard-educated carpenter" for example, people have attributed the property of being "idealistic," which is neither a typical property of "Harvard-educated" people or of "carpenters" (Hampton 1987). If concepts are sets of individuals, how do we conceive and reason about entirely novel concepts? Furthermore, it must be possible to *learn* new concepts we use, but how can we learn a concept if that means considering all individuals in the universe that satisfy that concept?

Fodor 1981 suggests provocatively that *all* concepts are innate and that humans are simply equipped with intensional functions which can tell us whether something is a member of a given concept. Of course, this radical proposition is intuitively rather implausible (Carey 2009), and psychologists have spent considerable effort trying to understand what could be the structure of this black-box that allows us to call a lion, a lion.

1.3.2 Psychological theories of concepts

Psychology always saw these issues with the view from philosophy and mainstream semantics as deeply damning to the project of understanding concepts. Psychologists were from the very beginning concerned with the plausibility of these notions as being implemented in human minds, and they were concerned with questions of learnability. Accordingly, their view has been more connected to the Russellian view of concepts as collections of properties, or *features*.

Prototype theory

One of the most influential theories of concepts in psychology is prototype theory (Rosch 1975). Under this view, concepts are represented as ideals to which things are compared. These prototypes are defined along dimensions called *features* which define constitutive properties of objects, abstract or concrete. For example, a prototypical lion has features like being large, predatory, having a mane or living on the savanna. We can tell whether something is a lion or not based on how close it fits this ideal; a female lion, despite not having a mane, still matches the other features very well.

This theory has been used to explain many different phenomena related to *typicality* in psychology. For example, a prototypical bird might be something like a robin, rather than a penguin. As such, people will think of robins before thinking of penguins when told to think of a bird, and they will categorise prototypical birds like robins as birds much faster than penguins in a variety of tasks (Rosch 1975).

Unfortunately, prototypes are problematic for compositionality. A classic critique of prototype theory is that a "pet fish" is something that is both a pet and a fish, but vitally a prototypical pet fish (e.g. a guppy) is neither a prototypical fish (e.g. cod, salmon), nor a prototypical pet (e.g. dog, cat), nor a combination of the features of a prototypical fish and prototypical pet (Osherson and Smith 1981; Fodor and Lepore 1996). In other words, if concepts are collections of features, we have little idea of how to combine the features of individual concepts into novel concepts (Fodor and Lepore 2002), despite many valiant attempts (Smith and Osherson 1984; Kamp and Partee 1995).

Another central problem with this class of theories is that it is unclear what the origin or nature of the features are. For example, take a plausible set of features that might be involved in the concept of a lion:

$$[\![\text{lion}]\!] = \langle +Predator, +HasAMane, +LivesOnTheSavanna, \ldots \rangle$$

It is impossible to place an individual in space along these featural dimensions without presupposing that we know the meaning of each feature. Unfortunately, some of these features seem best defined with respect to lions! For example, if we say that “lion” is partially defined by the presence of a mane, then we must define what a mane is: vitally without reference to a lion. Without a clear genesis of features (or an innate, finite set of features), we cannot use features without an infinite regress in meaning.

There are also many other theories of concepts in a similar tradition to prototype theory.

Exemplar theory

A common theory which is similar to prototype theory is exemplar theory (Nosofsky 1988). This theory replaces prototypes with *multiple* examples of each concept, or *exemplars*. Now, penguins, robins, etc. would all be different exemplars of birds to which new candidates could be compared. This is a much more flexible representation than traditional prototype theory. Indeed, some research has indicated that exemplar theory captures novel concept tasks more accurately than traditional prototype theory (Nilsson, Juslin, and Olsson 2008).

Conceptual spaces

There is also the approach of Gärdenfors 2000: Conceptual Spaces. This theory is in some sense, a generalisation of prototype or exemplar theory. A conceptual space is defined by having different features as dimensions. For example, some of the dimensions could correspond to height or hue for example.

Individual objects are then points in this space and *regions* of space define concepts. Rather than using specific points to determine concepts, we use *regions* of space to define a concept. These regions are defined by certain interpretable dimensions such as colour or size. Furthermore, Gardenförs introduces some new constraints on concepts. For example, he claims that regions must be convex (i.e. a straight line going between two points must be entirely contained inside the region) and connected (the region must be entirely one piece; there can’t be any “islands”).

Theory-theory or essentialism

Another approach that is quite popular in psychology is theory-theory. The idea is that we do not define concepts in terms of features alone, rather we develop theories for each concept.(Carey 2009) This is the result of classical experiments where even if a raccoon is modified to look exactly like a skunk, people would still consider it a raccoon (Keil 1989). If the original prototype theory were true, then the raccoon would be very similar to a skunk and so, should be considered a skunk, but people still consider it to be a raccoon.

The greatest successes of theory-theory approaches have been explaining concepts which define novel cognitive machinery such as the acquisition of counting (Carey 2009). This work is not focused on the kind of novel cognitive tools such as counting and rather focuses on simple concepts such as “shoe” or “lion”, so theory-theory’s accomplishments are less relevant to this work.

1.3.3 Natural language processing and concepts

Machine learning approaches concepts from a radically different way, primarily indifferent to their structure, seeking only *functional* representations that are learnt by optimisation algorithms. What I have reviewed in previous sections are all *symbolic* methods which either treat concepts as fundamentally abstract sets, or their features as abstract sets. Modern AI techniques are largely *subsymbolic* and eschew these kinds of explicit representations. Rather, representations emerge within the calculation of difficult-to-interpret neural networks which are trained on a given task.

In natural language processing (NLP), AI’s approach to language, this has had great success. We now have highly accurate machine translation systems, incredible text generation and impressive tools that can use language, such as Alexa. These systems do not represent words or concepts as bundles of features or abstract sets, but as *vectors* (Mikolov et al. 2013; Devlin et al. 2019). Unlike the features of prototype theories, the dimensions of these vectors do not have an explicit, underlying meaning. These rich representations excel at many language tasks and can predict some tricky semantic notions like analogy or similarity in humans (De Deyne, Perfors, and Navarro 2016; Mandera, Keuleers, and Brysbaert 2017).

Despite these representations’ effectiveness at many different types of tasks, they struggle with modelling some of the most basic inferences that people make. Instead, they use incredibly superficial heuristics and cannot even handle basic notions like negation correctly (McCoy, Pavlick, and Linzen 2019; Ettinger 2020). While it is intriguing that these machines are capable of understanding the similarity between concepts, they still seem to have a great deal of difficulty with *compositional* element of language. Indeed, some critiques of modern AI have focused on its profound difficulties handling the rich compositionality that is intrinsic to human language and thought (Marcus 2018).

1.3.4 Interdisciplinary theories of concepts

There has been limited interest in trying to combine these different approaches in an interdisciplinary fashion.

There have been some efforts to take word vectors seriously in linguistics, particularly due to their success at notions like similarity. Some attempts have tried to explicitly combine the compositional syntax of formal semantics with vector representations for all words, logical or otherwise (Baroni, Bernardi, and Zamparelli 2014). Others have looked at how these models can inform semantic theories about notions of similarity or polysemy (Boleda 2020). Conversely, others have looked at whether word-vectors learn theoretical contrasts predicted by formal semantics (Goodale 2022).

Natural language processing and psychology have also inspired each other. There have been accounts of the conjunction fallacy along the lines of Tversky and Kahneman's representativeness heuristic using word-vectors (Bhatia 2017). NLP models have also been proposed as models of concepts more generally (Bhatia and Aka 2022), despite these models difficulties with basic reasoning tasks or entailments.

In formal semantics, there has also been some work seeking inspiration from prototype theory or other psychological theories (P. Johnson-Laird 1982; Kamp and Partee 1995; Pustejovsky 1995; Mascarenhas 2014). For example, accounts of privative adjectives have been proposed on the basis of prototypical or conceptual representations (Del Pinal 2015; Guerrini and Mascarenhas 2019) although such accounts either have difficulty with recursive applications of privative adjectives (Martin 2019; Guerrini 2021) or implicitly assume feature-based representations. Generic statements have also been argued to be psychological "defaults" albeit without formalisation (Leslie 2007).

Chapter 2

Introduction to the tripartite view of concepts

The core idea of this work is that each concept, C, is defined as a triple: $\langle M_C, P_C, g_C \rangle$. M_C is a manifold of a concept which defines its extension. A manifold is a mathematical object from topology which locally looks like Euclidean space and has many nice properties which are appropriate for our conceptual representations. P_C is a probability distribution that allows us to sample exemplar points from M_C. Finally, g_C is a metric tensor: a mathematical object that allows us to define distances between points on M_C. Concepts are embedded in a general conceptual space, the $\mathscr{E}$-space. Concretely, we say that all individuals and concepts are embedded in the $\mathscr{E}$-space ($\mathscr{E}$ since it embeds objects which are traditionally of type e). The $\mathscr{E}$-space is represented as a high dimensional vector space, $\mathbb{R}^n$.

This is not an account of *how* concepts are acquired. Rather, the goal of this theory is to provide a minimum set of conceptual operations that should be available in order to engage in the kind of cognitive phenomena this work aims to address. With a more precise scope established, we can outline the theory and its broad contours before moving on to a formal definition in chapter 3.

2.1 Conceptual spaces without features

Simple categories such as “lion” have extensions that are *regions* of a high dimensional space. Unlike conceptual spaces (Gärdenfors 2000), it does *not* assume the dimensions of this space are interpretable or concretely definable.

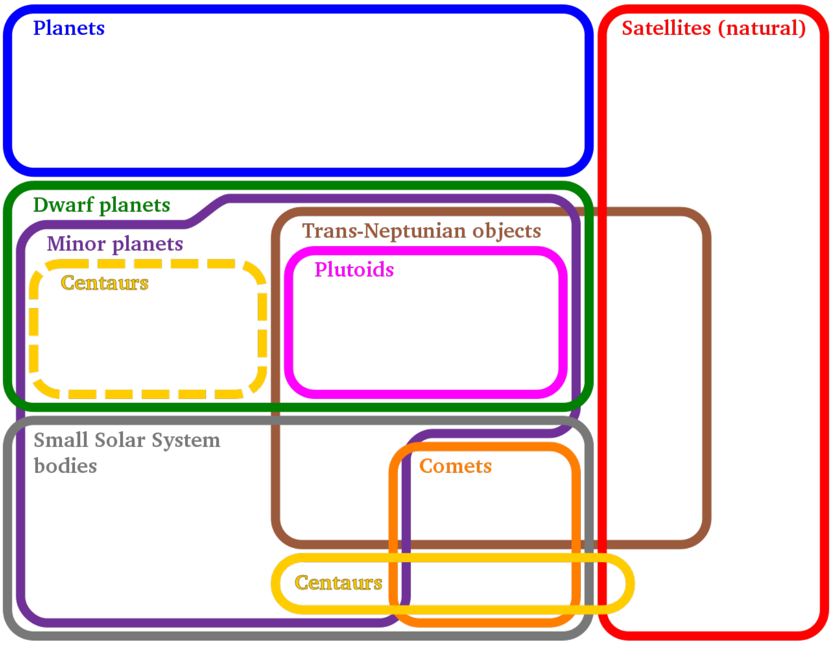


Figure 2.1: An example of a Euler diagram classifying different bodies in the Solar System. From Wikipedia

2.1.1 Conceptual regions

These regions could be considered like an *Euler diagram*, a classical tool for representing categories (See Figure 2.1 for an example). The dimensions of an Euler diagram have no inherent meaning at all; it's just a way of drawing arbitrary set-extensions graphically. While we would certainly have more than just 2 dimensions, the situation for this theory is the same: dimensions have no intrinsic meaning. The key *difference* is that the distance between points *does* relate to their similarity.

Simple concepts like "lion" are sets of points. We can, for the time being, treat individuals as simple points in this high-dimensional space. Ultimately individuals will be handled differently (c.f. Section 2.5)), but this simplification will be useful for now. A lion then, would simply be something which is in the region of space corresponding to lions.

This doesn't seem to give us much new power, but now, each individual has a rich representations *beyond* which sets they belong to.

2.1.2 Features are emergent, not foundational

These richer representations are *not* organised around coherent features or properties with an interpretable meaning. More radically, the claim is that features are *not* fundamental different from concepts, and that many of the features that are used to describe a concept are, in fact, simply other concepts.

Features are defined in many different ways across the literature, but what I am addressing is the idea that the underlying dimensions of conceptual space are interpretable or coherent.[1] That is, the dimensions are features such as "height", "weight", et cetera. Rather, I think the dimensions ought to be ineffable and lacking any inherent meaning.

There are numerous reasons to be sceptical of interpretable dimensions as a foundational part of concepts. The core problem behind interpretable dimensions is that they are ill-defined and define concepts in terms of other concepts. For example, take a plausible set of features that are involved in the classification of a lion:

$$[\![\text{lion}]\!] = \langle +Predator, +HasAMane, +LivesOnTheSavannah, \ldots \rangle$$

It is impossible to place an individual in space along these featural dimensions without knowing the meaning of each feature. Unfortunately, some of these features seem best defined with respect to lions! For example, if we say that "lion" is partially defined by the presence of a mane, then we must define what a mane is: vitally without reference to a lion. Without a clear structure for features (or an innate, finite set of features), we cannot use features without an infinite regress in meaning. Furthermore, it seems quite difficult to define every concept in terms of a *single* list of features.

The vast successes in NLP have been powered by representations of words with entirely arbitrary dimensions that have no clear meaning. Instead, meaning is diffused over the entire space, and it is the *relationships* between points that structures them. For example, two word-embedding spaces learnt on different corpora in different languages in the same semantic domain can be aligned (Conneau et al. 2018) to find word-to-word translation without any labeled data (c.f. Figure 2.2 for an illustration). For example, a mapping can be found between "gatto" and "cat" without knowing these words are translations of each other, simply based on the *shape* of the word-vectors. This is because the *relationships* between points define words, not their point in space per se. These spaces naturally have completely different dimensions, but nevertheless the spaces can be transformed from one to the other since they have the same basic "shape".

[1] Incidentally, in the machine-learning literature, the interpretable dimensions which are used as input for a classifier are typically called "features."

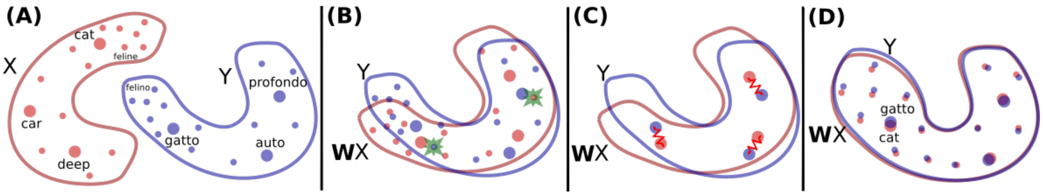


Figure 2.2: Figure reproduced from Conneau et al. 2018 which illustrates how word-embeddings trained on different languages can be aligned since the relationships between words remain analogous.

It is perfectly plausible for *some* dimensions to have an innate interpretable meaning, particularly those which might be innate to our cognition. Furthermore, it is extremely likely that linear subspaces of the space might self-organise around what seems like interpretable dimensions. Nonetheless, requiring dimensions to be *intrinsically* interpretable or coherent by themselves is untenable.

2.2 Extensions as manifolds

Our concepts have manifolds which carve out a region of the $\mathscr{E}$-space. “Region of space” is not a very precise definition; plenty of theories of concepts, including exemplar or prototype theory, have regions of space that correspond to given concept. The question is, how do we constrain these regions?

It’s important to note that these regions *overlap*; a lion, laying in the lion manifold *also* lies in the animal manifold. Prototype theory often draws distinctions between “basic-level” and “superordinate” categories (Rosch et al. 1976). This distinction is not necessary in this approach, as the hierarchical structure between concepts is implicitly encoded. For example, lions just happen to be a subset of animals.

2.2.1 Connectedness and convexity

A major contribution of Gärdenfors 2000 is the notions of *connectedness* and *convexity* for conceptual spaces. A concept is connected if and only if a path can be drawn between any two points in the concept, e.g. there are no separate “islands”. Conversely, a concept is convex if a *straight* line can be drawn between any two points in the concept where the entire line remains in the concept. There is some evidence that people prefer to learn connected concepts over disconnected ones, but we are capable of learning disconnected concepts as well. (Xu and Tenenbaum 2007; Chemla, Buccola, and Dautriche 2019). This tendency has also been observed in non-human primates (Chemla, Dautriche, et al. 2019), These kinds of inductive biases might be useful, but they needn't be a hard constraint. A considerably weaker desideratum for these regions might be that they are *locally* connected and convex. In other words, given any point in a concept, can we go in an infinitesimally small length in any direction and remain in the same concept? This will constrain us considerably in the shapes we can draw, but give us considerably more expressive power than convexity. A careful reader might notice that this seemingly arbitrary definition, is in fact an informal definition for *a manifold*.

2.2.2 Manifolds oversimplified

While many linguists are well-versed in discrete mathematics and set theory, other fields of mathematics such as topology are understudied. In particular, topology and differential geometry provide us with the right tool to properly discuss similarity: the *manifold*. Our "regions", are in fact, manifolds.

Informally, a manifold is a space which locally resembles Euclidean space. For example, the Earth seems flat at any given point, but is actually a sphere: a 2D manifold embedded in 3D space. It is 2D because it locally looks like a *plane*.

Manifolds have been applied in countless scientific disciplines. For example, in computational neuroscience there is increasing interest in modelling neural activity with manifolds. Concretely, one defines a space where each dimension corresponds to the activity of a different neuron. The point representing the population of neurons at a given time will typically be constrained to a specific manifold with a much lower dimensionality than the number of neurons. Most intriguingly, specific actions or objects are then associated with different submanifolds embedded in a larger manifold (Gallego et al. 2017; DiCarlo, Zoccolan, and Rust 2012; Saxena and Cunningham 2019).

A manifold, M is a tuple $\langle X, \mathscr{T} \rangle$. X is a set of points and $\mathscr{T}$ is a topological structure that defines "closeness" between points. This defines the open sets that make up that topology. We can think of open-sets as being a generalisation of infinitesimally close points or of open intervals, e.g. (0,1). For example, given a point x, we can define an open ball around x of radius r:

$$\mathfrak{B}(x,r) = \{y \in \mathbb{R}^n | d(x,y) < r\}$$

In other words, all points that are a distance of less than r away. This ball defines an open set.

More concretely, a topology is any tuple, $\langle X, \mathscr{T} \rangle$ such that $\mathscr{T}$ meets the following conditions:

i. X and $\varnothing$ are in $\mathscr{T}$.

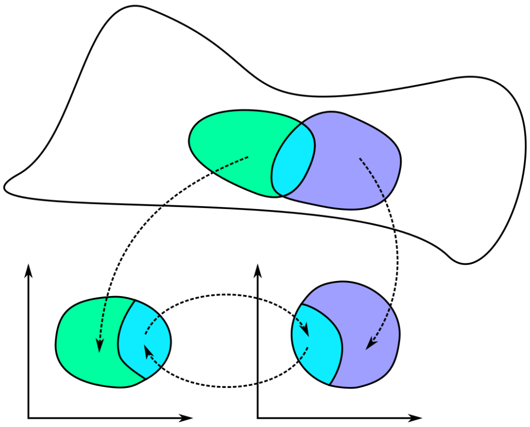


Figure 2.3: An example of a differentiable atlas for an arbitrary manifold. The top image represents the manifold, where the green area is U, purple is V and cyan is $U \cap V$. The arrows are the different functions which map from U or V to $\mathbb{R}^n$. All of these functions should be differentiable. Image courtesy of Wikipedia.

ii. Any union of members of $\mathscr{T}$ is in $\mathscr{T}$.

iii. Any intersection of a finite number of elements of $\mathscr{T}$ is in $\mathscr{T}$.

This means that all the subsets of $\mathscr{T}$ are open sets.

Not all topologies are manifolds, but manifolds are special kind of topological space. A manifold has the condition that each point locally resembles a Euclidean space. So, at any given point, it looks like we are on an n-dimensional plane if we are a n-dimensional manifold. For example, the surface of the Earth locally looks like a 2D plane, so it is a 2D manifold, even if we are embedded in 3D space.

Furthermore, if we wanted to be able to measure distances on our manifold, it must be a *differentiable* manifold. This is necessary to define things like *lengths* over the manifold. For an open set $U \in \mathscr{T}$, there exists a *chart*, (U, f) such that $f : U \to \mathbb{R}^n$.[2]

A chart is a way to define a coordinate system for the part of the manifold covered by U. For example, a map of Earth defines a chart for the earth, even though there will be distortion near the poles. f must be a bijective function such that f and its inverse, f^{-1} are continuous. If f^{-1} is differentiable, then we can use it to do calculus on M and thus measure lengths on M.

[2] Normally this mapping is referred to as φ not f, but to avoid confusion with propositions, I have written it as f.

However, we might need more than one chart for all the open sets of the manifold, and those charts might share points between them. In order to ensure that we can properly differentiate over M, we need to make sure that we can smoothly transition from one chart to another. Concretely, say that we have two charts, (U, f_1), and (V, f_2) where $U \cap V \neq \varnothing$. Then, we must ensure that $f_2 \circ f_1^{-1}$ is differentiable so that we can smoothly transition from one chart to the next. While this remains a bit complicated, Figure 2.3 shows such an example. If we can find a set of charts that covers all of M and that has these smooth transitions, then we have a *differentiable atlas*, and thus, M is differentiable and we can safely define distance along it.

2.3 Similarity

2.3.1 Metrics and similarity

Once we've organised our points in $\mathscr{E}$ into manifolds, we also want to see how similar two points are. To do so, we add a *metric* to this space. A metric is a kind of *distance* function. There are many possible metric functions, the classic example is normal Euclidean distance:

$$d(\vec{x}, \vec{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

Specifically, a metric, $d : M \times M \Rightarrow \mathbb{R}$ is a function over a set M which meets the following properties.

i. **Identity of indiscernibles:** $d(x, y) = 0 \Leftrightarrow x = y$

ii. **Symmetry:** $d(x, y) = d(y, x)$

iii. **Triangle Inequality:** $d(x, z) \leq d(x, y) + d(y, z)$

The use of metric spaces for similarity has an extremely long history in psychology (Shepard 1962).

Often, this metric is assumed to be Euclidean, but this is not necessarily a good thing. For example, humans are known to often classify things into hierarchical categories. "Furniture" is a relatively broad category, "chairs" less so, and "thrones" are a very particular kind of chair. Often, these categorisations are represented as trees, but trees are difficult to embed in Euclidean space without considerable distortion. Metrics in *hyperbolic* space, do not suffer from this problem and are, in fact, the continuous analogy to trees (see Figure 2.4 for a visualisation). This property of hyperbolic distance metrics has been exploited

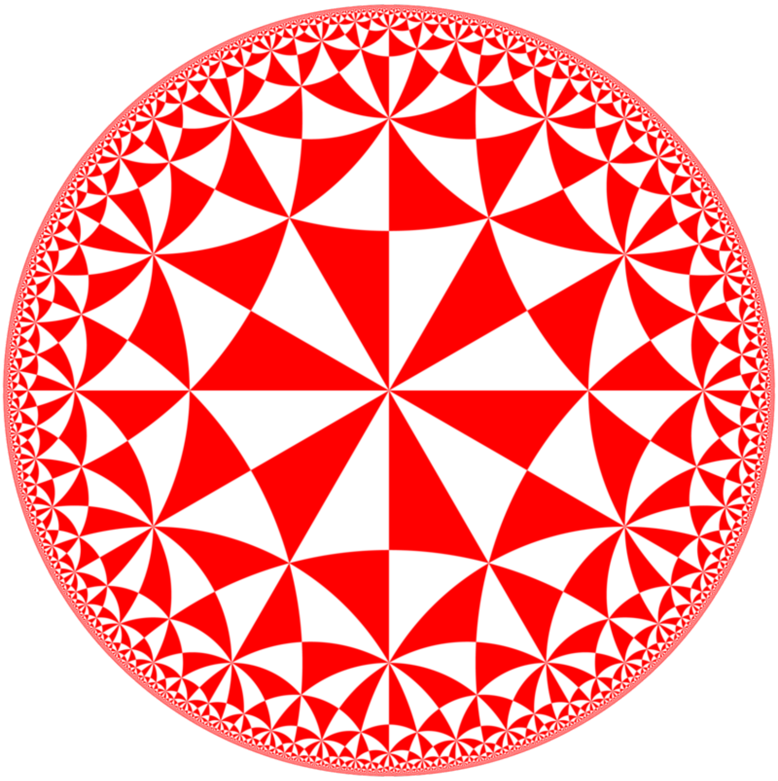


Figure 2.4: A visualisation of hyperbolic space. Each tile edge is the same length according to the hyperbolic distance metric. All points are relatively close to the centre, but can be very far from one another. Notably, M.C. Escher's Circle Limit III uses the same tiling as this example. Figure from Wikipedia.

recently in machine learning, particularly for representing words or human categories (Nickel and Kiela 2017; Ganea, Becigneul, and Hofmann 2018; Suzuki, Takahama, and Onoda 2019).

For similarity, we use the Universal Law of Generalisation proposed by Shepard 1987 shown in equation 2.1. Concretely, this function maps distances from $[0,\infty)$ to similarities in range $[1,0)$. Very small distances are close to 1 (indeed, $e^0 = 1$), so things which are close together are very similar. Conversely, as the distance increases, the similarity decreases exponentially. We say that $\vec{x} \approx \vec{y}$ is a function with range $(0,1]$ which represents the degree to which x is similar to y. Note that we write $\overrightarrow{\mathrm{x}_{\mathscr{E}}}$ to indicate that the variable is a point in $\mathscr{E}$.

$$\overrightarrow{\mathrm{x}_{\mathscr{E}}} \approx \overrightarrow{\mathrm{y}_{\mathscr{E}}} = e^{-d(\overrightarrow{\mathrm{x}_{\mathscr{E}}},\overrightarrow{\mathrm{y}_{\mathscr{E}}})} \tag{2.1}$$

2.3.2 Similarity is always relative

The symmetric requirement (ii) has been criticised (Tversky 1977). For example, people are more willing to say (6a) than (6b).

(6) a. The son is like the father.

b. The father is like the son.

While this shows that a metric cannot naïvely represent the complex meaning of compositional sentences like those in (6), it does not mean that the underlying space need not be metric, nor that conceptual spaces should be abandoned. The solution to this proposed by Tversky is to look at the shared features of two objects, but we cannot resort to features in our system. Another tempting solution to propose that two things are similar if they share predicates. That too, is marred with problems, as pointed out by Goodman 1972, since for any two objects there are an *infinite* number of shared predicates between them. Goodman thinks for this reason, among others, that we should abandon similarity as an explanation for philosophical phenomena. Linguistics and psychology, unfortunately, will still have to grapple with the "insidious" thing that is similarity.

Context-sensitivity

Beyond its violation of reflexivity, similarity is also deeply sensitive to context; two objects are compared differently depending on the situation. Goodman cites the following example (Goodman 1972, p. 445):

> [S]uppose we have three glasses, the first two filled with colorless liquid, the third with a bright red liquid. I might be likely to say the

> first two are more like each other than either is like the third. But it happens that the first glass is filled with water and the third with water colored by a drop of vegetable dye, while the second is filled with hydro-chloric acid—and I am thirsty

In terms of colour, the glasses can be grouped in one sense, or can be grouped in terms of the drink inside them. Evidently, this scenario is impossible to handle with a single similarity measure. Tversky himself handles this kind of context-sensitivity by weighting features differently depending on context and the objects being compared.

Gardenförs deals with this context-sensitivity in a very natural way; objects are compatible with different conceptual spaces and can be compared in different ways (Gärdenfors 2000). For example, in Goodman's case, we are comparing the colours which will have different distances in the colour-space than the objects would have in say, gustatory-space (given hydrochloric acid's rather particular taste). Likewise, this contextual-sensitivity can deal with symmetry; the order of presentation in a sentence may very well affect the way in which things are compared.

2.3.3 Similarity along a manifold

To handle contextual-similarity, we simply look at distances *along* a manifold. To get intuition about this kind of distance, consider the distance between Paris and New York. If we took the direct Euclidean distance between them, we would travel through the mantle of the earth: a horrid flight-plan. Rather, we follow the surface of the Earth to determine their distance, curving through space to stay on the Earth's surface. Likewise, we could define a distance *within* any conceptual manifold rather than through $\mathscr{E}$-space itself, even though the manifold lies within $\mathscr{E}$-space. Concretely, when we compare two objects, we are always comparing them *within* a relevant domain.

For example, consider two individuals, John and Jim. They are very similar people; they're both young brown-haired men with fair-skin with similar hobbies and interests. They're also both guitarists. As guitarists, however, they couldn't be more different. John plays precisely and methodically on an acoustic guitar; Jim plays fast and rough on an electric guitar, making heavy use of the whammy bar.

(7) As a person, John is like Jim.

(8) As a guitarist, John is not like Jim.

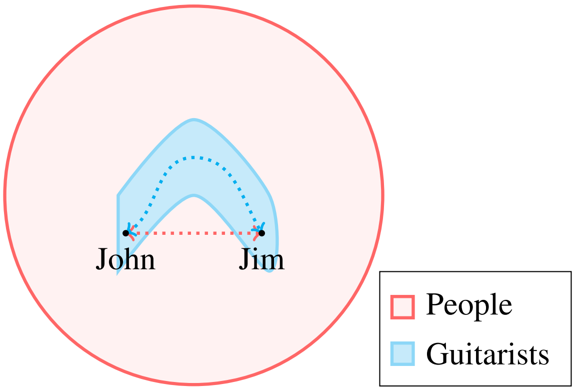


Figure 2.5: A diagram showing the different paths that would define the distances between Jim and John depending on if we compare them as guitarists or as musicians, as in (7) and (8).

We can say that both (7) and (8) are true in this scenario. Vitally, this turns similarity from a two place relation to a three place one, as shown in equation 2.2. We have our two objects, x and y, and the category, C within which we are comparing them. Vitally, C must be a set that contains both x and y, otherwise the distances between them will be undefined.

$$\overrightarrow{x_{\mathscr{E}}} \underset{C}{\approx} \overrightarrow{y_{\mathscr{E}}} = e^{-d_C(\overrightarrow{x_{\mathscr{E}}},\overrightarrow{y_{\mathscr{E}}})} \tag{2.2}$$

$d_C(x,y)$ is just the length of the shortest path along the manifold from x to y. To measures length on a manifold, we endow our manifolds with a metric tensor, g_C. To visualise what this might look like for (7) and (8), see Figure 2.5.

Our context sensitivity for similarity comes from the d_C we use. This will be vital for various linguistic applications, particularly privative adjectives as discussed in Chapter 5.

2.3.4 The metric tensor: g

The metric tensor, g defines the distance to the infinitesimally small neighbours of each point on M.[3] Concretely, g is a tensor field; for each point of the manifold, there is an associated matrix which defines distances around that point. Importantly, g is not an intrinsic part of the manifold; it must be chosen.

We can think of g as kind of scale bar on a map. Given coordinates which describe the position of a point on a manifold, g tells us how to adjust distances

[3] Technically g defines an inner-product on the tangent space of M which then can be used to define the distance between a point and its neighbours.

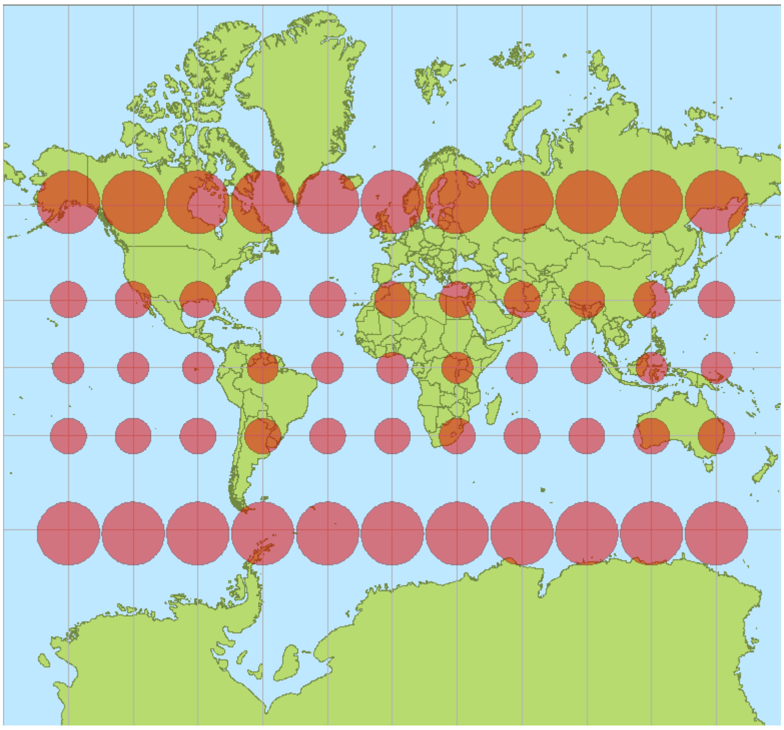


Figure 2.6: An illustration of the distortion on the Mercator projection from Wikipedia. g can be understood as roughly the relative sizes of those dots.

around that point. So, for example, on a 1mm:1m map, g would tell us to multiply all lengths by 1000. Our g is more general however, since the distortion changes depending on where we are on the manifold. For example, the scale of the Mercator projection gets larger near the Earth's poles, as shown in Figure 2.6.

For normal Euclidean space with Cartesian coordinates in $\mathbb{R}^3$, the point, $\langle x,y,z\rangle$ has neighbouring points $\langle x+dx, y+dy, z+dz\rangle$. To measure the distance, between $\langle x,y,z\rangle$ to its neighbours, we can use the following equation:

$$ds = \sqrt{dx^2 + dy^2 + dz^2} \tag{2.3}$$

We can make this more general with g:

$$ds = \sqrt{g_{00}dx_0dx_0 + g_{10}dx_1dx_0 + \ldots} = \sqrt{\sum_{i=0}^{n}\sum_{j=0}^{n} g_{ij}dx_idx_j} \tag{2.4}$$

The metric tensor for Euclidean distance in $\mathbb{R}^3$ is the identity matrix:

$$g_{euclid} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.5}$$

If we plug g_{euclid} into 2.4, it reduces to 2.3, but we could use different matrices instead for different metrics. Crucially, this matrix changes depending on the point we measure at; it's not necessarily the same for each point.

Distance on a manifold

Once we have g_C and M_c for a given concept, C, we can define the distance, d_C, between two points in M_c as the length of the shortest continuous curve which is entirely on the manifold. Any of our manifolds can be turned into a metric space using this definition (Lee 1997). Computing this distance can be remarkably complicated, but there are approximations in polynomial time (Crane et al. 2020). Note that g_C allows us to measure lengths, but the shape of the manifold determines what curves are possible and thus the length of the shortest curve is.

Note however, that if the manifold is *disconnected* (i.e. there are different "islands" that make up the manifold), the distance between those points will be undefined since there is no path that connects them. Thus we need Gardenförs's connectedness restraint if we want to be able to compare any two members of a concept, C with d_C. Alternatively, we could simply state that the distance between two points that are disconnected is infinite, and so the similarity, $\overrightarrow{\mathrm{x}_{\mathscr{E}}} \underset{C}{\approx} \overrightarrow{\mathrm{y}_{\mathscr{E}}}$ will be 0, since $\lim_{x \to \infty} e^x = 0$

Convexity, however, we get for free if we are connected. The principal reason that Gardenförs introduces convexity is so that *betweenness* is well-defined. Essentially, it states that if x and y are both members of a concept, then something between them ought also to be a member of the concept. Since d_C is determined by curves on the manifold of C, any point along that curve will, by definition, be on the manifold. This trivially satisfies Gardenförs's motivation for convexity.[4]

2.3.5 Metaphorical use of predicates

Now that we have characterised two of three ingredients in a concept, we can go over some specific details. For example, how do we tell if a point is a member of a concept? Literal predication is done with the following equation, where C is a given concept:

$$C(\overrightarrow{\mathrm{x}_{\mathscr{E}}}) = \begin{cases} 1 & \text{if } \overrightarrow{\mathrm{x}_{\mathscr{E}}} \in M_C \\ 0 & \text{if } \overrightarrow{\mathrm{x}_{\mathscr{E}}} \notin M_C \end{cases} \tag{2.6}$$

[4] Indeed, Gardenförs rightly points out that certain domains such as colour would require a different metric to satisfy convexity (Gärdenfors 2000). However, he views this metric as something that is born out of the interpretable features or dimensions of a given conceptual space, rather than a *lexical* part of a concept, independent of the dimensions it is embedded in.

But, we can also use a predicate metaphorically. For example, if a man was particularly adept at fighting, we might describe him as a lion, since they are both brave or powerful or so on. Furthermore, we constantly refer to objects by the name of another kind of object without even thinking of it as metaphor; for example, it is entirely natural to call a statue of a lion, a lion. (Franks 1995).

Contextual similarity gives us a way to address this. The man who is a lion in terms of his fighting skill, would be close to being a lion along the distance metric defined by "fighter". We can notate this with a special operator, $(A \upharpoonright B)(\overrightarrow{\mathrm{x}_{\mathscr{E}}})$, which returns the distance of $\overrightarrow{\mathrm{x}_{\mathscr{E}}}$ to the A manifold with respect to the B concept.

To do this, we use $d_B(\overrightarrow{\mathrm{x}_{\mathscr{E}}}, M_A)$ to measure the distance from $\overrightarrow{\mathrm{x}_{\mathscr{E}}}$ to all points on M_A. From all this distances, we choose the closest and see how close $\overrightarrow{\mathrm{x}_{\mathscr{E}}}$ is to it. In other words, how far is $\overrightarrow{\mathrm{x}_{\mathscr{E}}}$ from being a member of A? If it's very close to being a member (using the distance metric d_B), then we say that $\overrightarrow{\mathrm{x}_{\mathscr{E}}}$ is an A metaphorically. If it is very far from being a member, then it is not.

$$(A \upharpoonright B)(\overrightarrow{\mathrm{x}_{\mathscr{E}}}) = \max_{\overrightarrow{\mathrm{y}_{\mathscr{E}}} \in M_A} \left(e^{-d_B(\overrightarrow{\mathrm{x}_{\mathscr{E}}}, \overrightarrow{\mathrm{y}_{\mathscr{E}}})} \right) \tag{2.7}$$

Say John is a lion when it comes to fighting. John would have a very high value for $(Lion \upharpoonright Fighter)$ because along the fighter-concept, he is close to being a lion. He could still remain far from being a lion in many other categories, for example, in terms of appearance (or perhaps personality, if he is not much of a lion in his social life).

There could be some primitive concepts which are used as a default distance. For example, the physical appearance of a thing is a very typical way to compare two objects: this would be the manifold of all things that have a physical appearance. The distance metric for that manifold would then describe how similar two things are *qua* visual appearance.

2.4 Typicality

While now we are able to describe the extension of a concept and see how points are similar; there is still a missing ingredient. We still have no notion of *typicality*.

Typicality is the last thing which is necessary to correctly model a huge number of phenomena in human reasoning (Tversky and Kahneman 1974; Sablé-Meyer and Mascarenhas 2021) and many different linguistic ones as well, particularly things like generics (R. v. Rooij and Schulz 2020; Cohen 1999; Leslie 2007).

It might be tempting to say that typicality is the distance from the center of the manifold. This would be roughly equivalent to prototype theory where a single ideal point is considered the most typical and to which things are compared. There is not, however, a clear definition of *center* on a manifold. For example, we could say that the centre, c, of the manifold, M, is the point which is as close as possible to all other points on the manifold. Let $x_1, x_2, \ldots, x_n$ be a set of equally distributed points on the manifold M.

$$c = \operatorname*{argmin}_{p \in M} \sum_{i=1}^{n} d_M\left(p, x_i\right) \tag{2.8}$$

Equation 2.8 shows such a calculation, known as the *Karcher mean*. The problem, of course, is that this does not always pick out a unique point! Consider a sphere: any point on that sphere will be a Karcher mean by this definition.

Instead, we can endow our concepts with that final bit of structure: a probability distribution over the manifold. This probability *does not* refer to the real world probabilities *nor* beliefs about probabilities: rather it is simply the probability of sampling a given point from a manifold. Sampling from this probability distribution allows us to *quantify* and then implicitly causes typicality.

2.4.1 Quantifying with typicality

To interpret (9), we need to ensure that that the manifold of lions, L, lies entirely within the set of things with tails, T. [5]

(9) All lions have a tail.

We could check whether the lion-points are a subset of the tail-having-points, but since we have uncountably many lion-points, this would require evaluating a potentially intractable integral. A trivial approximation would be to make a *finite* set by randomly sampling from L and checking if all those points are in T. The points sampled are referred to as *exemplars*, but unlike exemplar theory, they are not stored in memory. Let $\mathsf{Ex}(L, n)$ be a random distribution which samples n exemplars independently from L according to the probability distribution defined on L. Then, we simply quantify over a random variable, $X \sim \mathsf{Ex}(L, n)$ rather than over L directly. For brevity in equations, we simply write $\mathsf{Ex}(n, P)$ rather than $X \sim \mathsf{Ex}(n, P)$.

$$\begin{aligned} [\![\text{All lions have a tail}]\!] &= \forall (x \in L) T(x) \\ &= \min_{\overrightarrow{\mathtt{x}_{\mathscr{E}}} \,\in\, \mathsf{Ex}(L,n)} T(\overrightarrow{\mathtt{x}_{\mathscr{E}}}) \end{aligned}$$

[5] Actual quantifiers in natural language are, of course, more complicated than this, but we still need a universal quantifier in our system.

Recall that $T(\overrightarrow{\mathrm{x}_{\mathscr{E}}}) = 0$ if $\overrightarrow{\mathrm{x}_{\mathscr{E}}} \notin T$ and $T(\overrightarrow{\mathrm{x}_{\mathscr{E}}}) = 1$ if $\overrightarrow{\mathrm{x}_{\mathscr{E}}} \in T$, so the minimum will be 1 if and only if every exemplar has a tail, and 0 if any does not. If n is sufficiently small, this will sometimes lead to false judgements and we say that n varies with effort or with the looseness of the conversation.

Indeed, people will sometimes agree to universally quantified statements that are trivially false but agree with their extension, like (10) (Leslie, Khemlani, and Glucksberg 2011).

(10) All lions have manes.

This is because we sample typical exemplars much more that atypical ones. Concretely, for a concept like "birds", the probability mass would be much greater around small songbirds like robins rather than atypical birds like ostriches or penguins. Of course, sampling frequency is not determined solely by real-world frequency nor beliefs about real-world frequency.

Concepts could share the same manifold, yet have different probability distributions. This is because the same extension can have different typical exemplars. For example, the extensions of "drinks" and "beverages" might be identical, but a typical drink might be something like water or beer, whereas a typical beverage might be something like Coca-Cola or tea. So, when we sample from "drinks" we have a different probability distribution than "beverages", even if they exist on identical manifolds. This will be key to dealing with Frege's puzzle later on in Section 2.5.2.

2.5 Uncertain individuals

Thus far, I have been treating individuals as specific points in $\mathscr{E}$ for didactic purposes. Instead, individuals are better modeled as *individual concepts* which are triples just like category concepts like "lion".

The reason we cannot model individuals as points is because we are not 100% certain about every single fact about an individual. If all individuals are single points in $\mathscr{E}$-space, just by virtue of knowing an individual exists, we automatically know everything about it. We know precisely which manifolds it lies on and which is does not and therefore we are completely convinced about which predicates are true for it. The idea of automatically being certain about every fact about an individual simply by representing that individual seems hard to swallow.

By treating individuals as concepts, we can allow for *uncertainty*. We now have two kinds of concepts which are represented identically. Category concepts are concepts like "lion", where the manifold corresponds to the extension of possible members. Conversely, for individual concepts like "Simba", the manifold corresponds to a set of possible candidates. The probability mass corresponds to our best guesses for an individual. Conversely, each point on the manifold represents a different candidate which is compatible with the information we have learnt about that individual.

When we shrink the individual candidate manifold, we reduce the space of possible candidate individuals. Conversely, shrinking the manifold of a category concept changes the extension of that concept. So, if we aim to minimise our uncertainty, we would want to make our individual concepts as small as possible.

2.5.1 Predication with uncertain individuals

Predication was a simple matter when individuals were points. For individual concepts, we simply *quantify* over our candidate points instead of assuming that a given point is the complete individual concept. However, it's not clear we should use either universal quantification (since then we are restricted to facts we know with 100% certainty) nor existential quantification (since then we are too lenient with predication). Instead, we could simply look whether the individual "tends" to have a certain predicate true of it.

Predication for individual concepts is the *expected truth value*. Given a individual concept, C, and a predicate, Q, we can define this more precisely, where $\overrightarrow{\mathtt{x}_{\mathscr{E}}}$ is sampled from the probability distribution of C. Note that $\mathbb{E}[X]$ corresponds to the expected value of X: the average result we expect to get from a random variable, X.

$$Q(C) = \mathbb{E}\left[Q(\overrightarrow{\mathtt{x}_{\mathscr{E}}})\right] \tag{2.9}$$

We use the same sampling procedure as we did in Section 2.4.1 to avoid computing a potentially intractable integral and by the Law of Large Numbers, as n increases we will approach the actual expected value. Concretely then, we define simple predication with equation 2.10

$$Q(M) = \frac{1}{n} \sum_{\overrightarrow{\mathtt{x}_{\mathscr{E}}} \in \mathsf{Ex}(M,n)} Q(\overrightarrow{\mathtt{x}_{\mathscr{E}}}) \tag{2.10}$$

Note that we can notate this by introducing a new quantifier, μ. This quantifier, μ, simply corresponds to the *mean* truth value, as made clear by equation 2.11. The shorthand $Q(M)$ is easier to read, but explicitly showing μ is very useful later on.

$$(\mu \overrightarrow{\mathtt{x}_{\mathscr{E}}} \in M)\, Q(\overrightarrow{\mathtt{x}_{\mathscr{E}}}) = \frac{1}{n} \sum_{\overrightarrow{\mathtt{x}_{\mathscr{E}}} \in \mathsf{Ex}(M,n)} Q(\overrightarrow{\mathtt{x}_{\mathscr{E}}}) \tag{2.11}$$

An example interpretation of a sentence like "John runs" could be:

$$\begin{aligned}
[\![\text{John runs}]\!] &= [\![\text{John}]\!]\left([\![\text{runs}]\!]\right)\\
&= \left(\lambda P.\left(\mu \overrightarrow{\mathtt{j}_{\mathscr{E}}} \in John\right) P(\overrightarrow{\mathtt{j}_{\mathscr{E}}})\right)\left([\![\text{runs}]\!]\right)\\
&= \left(\mu \overrightarrow{\mathtt{j}_{\mathscr{E}}} \in John\right)[\![\text{runs}]\!]\left(\overrightarrow{\mathtt{j}_{\mathscr{E}}}\right)\\
&= \frac{1}{n}\sum_{\overrightarrow{\mathtt{j}_{\mathscr{E}}} \in \mathsf{Ex}(John,n)} [\![\text{runs}]\!]\left(\overrightarrow{\mathtt{j}_{\mathscr{E}}}\right)
\end{aligned}$$

These values are *continuous*; statements are not simply true or false. This will naturally have implications for the logic which are addressed in Section 2.6. Note that while I illustrate the compositional nature of the system with a lambda calculus, I am not providing a full Monatgue grammar.

2.5.2 Frege's puzzle

We can now address a classic problem in philosophy of language from Frege (Frege 1892). These identity statements seem trivially true:

(11) a. John Smith is John Smith.

b. My cat is my cat.

c. Three is three.

None of the examples in (11) seem particularly interesting and each one is self-evidently true. Conversely, the following examples all seem much more meaningful:

(12) a. Cicero is Tully.

b. $e^{i\pi}+1$ is 0.

c. Superman is Clark Kent.

The problem is that if things like the examples in (12) are identical, then we should be able to substitute the different terms without changing the meaning. However, "Cicero is Cicero" and "Cicero is Tully" are different in meaning, even if "Cicero" refers to the same thing that "Tully" does. Frege suggested that this should be dealt with by adopting a secondary level of meaning beyond reference, *sense*. So, while all of these examples have denote the same objects, they do not necessarily have the same sense in (12).

This is an internalist approach so we are not considered with external reference. Nevertheless, we still need a way of saying when two things are "equal". A natural solution is when they both have the same manifold:

$$[\![=]\!] = \lambda X.\lambda Y.(M_X = M_Y) \tag{2.12}$$

This does not refer to equality in philosophical sense, but rather the equative construction in natural language (e.g. "is" in sentences like (11) and (12)).

Let's apply this to the classic scenario in (12a) We have two individual concepts, C and T for Cicero and Tully respectively. As a novice of Classical Antiquity, I may not know Cicero and Tully are the same individual, and wrongly believe they are different writers. As such, C and T would pick out different parts of the $\mathscr{E}$-space. If I then learnt that Cicero *is* Tully, I would then need to make C and T have the *same* manifold. They may still have different probability distributions or g_M however!

As a result, different predicates are true of each equal individuals concept. Consider Superman and Clark Kent: Clark Kent *qua* Clark Kent does not fly, yet Superman does fly.

(13) Superman flies but Clark Kent does not fly.

(13) produces a greater problem than Frege's original puzzle since now the flying predicate only applies to Superman and not Clark Kent, despite them being the same individual. In other words, when sampling from $P_{ClarkKent}$, Superman does not fly, whereas sampling from $P_{Superman}$, he does.

Since every point of Clark Kent lies within the Superman individual concept, those Clark Kent candidates are also potential Supermans and vice versa. Sense for the system, then comes from the relative likelihood of a given candidates being activated contextually. This also the exact same story we gave for synonyms with different connotations such as "drinks" and "beverages" in Section 2.4.

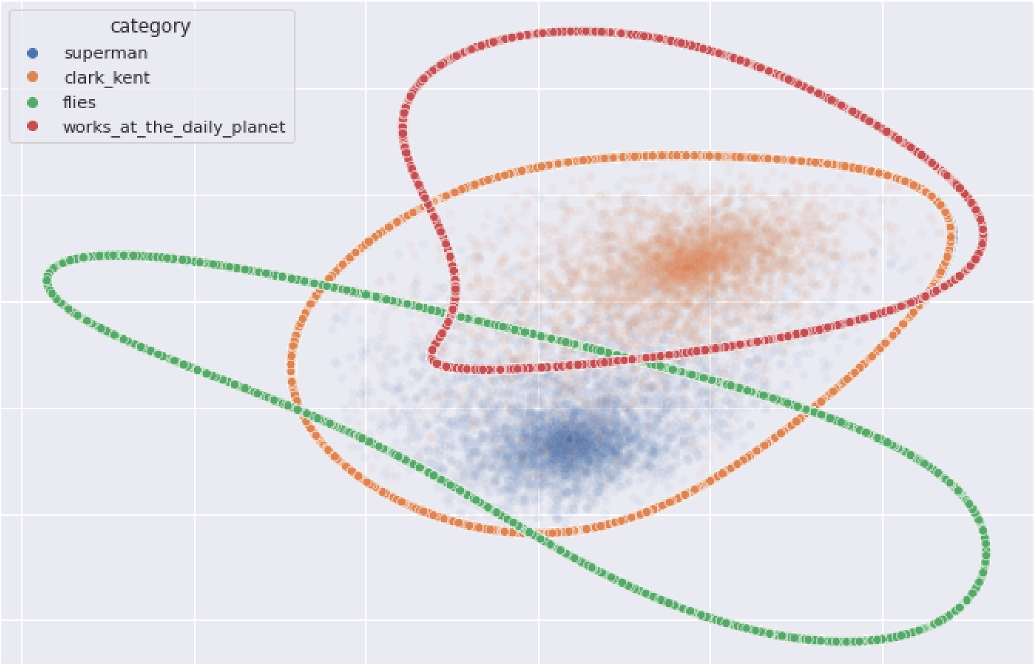


Figure 2.7: An example model that describes the scenario for Superman and Clark Kent. Note that the orange line defines the boundary of both the Superman and Clark Kent manifold. For a specification of how models like this can be learnt computationally, look at Section 3.6.

2.6 A continuous logic

Our truth-values are continuous. This is a kind of fuzzy logic (Zadeh 1965) where truth values range from 0 to 1 and the standard connectives $\vee, \wedge$ and $\neg$ are $\max(a,b), \min(a,b)$ and $1-x$ respectively. While this logic has seen considerable interest from both engineers and logicians, linguists have been considerably less enthusiastic. Following classical arguments from Fine 1975 and Kamp 1975, most linguists have been dissuaded from fuzzy logics. However, as Sauerland 2011 notes, this criticism may not be entirely well motivated, despite the problems with classical fuzzy logic since there have been innovations in the approach.

Specifically, the criticism leveled by Kamp and Fine came from the fact that $A \wedge \neg A$ is not necessarily 0 under a fuzzy logic. Consider proposition φ where the truth-value is 0.5, $\neg\varphi$ is therefore $1-0.5$, thus $\varphi \wedge \neg\varphi$ has a truth value of 0.5. This seems wholly unsatisfactory for most linguists, myself included. However, it seems in certain pragmatic contexts, people *do* prefer statements of the form $\varphi \wedge \neg\varphi$ to *both* φ and $\neg\varphi$ (Sauerland 2011).

Luckily, the way that our system is set up, we evade such scenarios because the individual points do not have continuous values for simple predicates. Concretely, every point is either on or not on a manifold, an individual point does not have a truth value. Thus, as long as we raise our new μ quantifier above all of the predicates, then we evade Kamp's scenario completely. For example, let's say that we have an individual, John, who may or may not be a musician. His probability distribution and manifold are arranged in such a way that we say that "John is a musician" has a truth value of 0.5. It's important to note that this truth value should not be understood as it is 50% true that John is a musician. Instead, it means something like one is 50% sure that John is a musician. While this bears a superficial resemblance to probability, the axioms of probability are violated by fuzzy logic (For example, $P(a \cap b) \neq \min(P(a), P(b))$).

$$\begin{aligned}
[\![\text{John is a musician}]\!] &= (\mu \overrightarrow{\mathtt{x}_{\mathscr{E}}} \in John)\, M(\overrightarrow{\mathtt{x}_{\mathscr{E}}}) \\
&= \frac{1}{n} \sum_{\overrightarrow{\mathtt{x}_{\mathscr{E}}} \in \mathsf{Ex}(John,n)} M(\overrightarrow{\mathtt{x}_{\mathscr{E}}}) \\
&= 0.5
\end{aligned}$$

Clearly "John is not a musician" has a truth value of 0.5:

$$\begin{aligned}
[\![\text{John is not a musician}]\!] &= \neg\, (\mu \overrightarrow{\mathtt{x}_{\mathscr{E}}} \in John)\, M(\overrightarrow{\mathtt{x}_{\mathscr{E}}}) \\
&= 1 - \left(\frac{1}{n} \sum_{\overrightarrow{\mathtt{x}_{\mathscr{E}}} \in \mathsf{Ex}(John,n)} M(\overrightarrow{\mathtt{x}_{\mathscr{E}}}) \right) \\
&= 1 - 0.5 \\
&= 0.5
\end{aligned}$$

The puzzle lies with the statement "John is not a musician and is a musician". Naïvely applying the compositional power of our connectives will get us the following (problematic) result:

$$\begin{aligned}
[\![\text{John is a musician and is not a musician}]\!] &= ((\mu \overrightarrow{\mathtt{x}_{\mathscr{E}}} \in John)\, M(\overrightarrow{\mathtt{x}_{\mathscr{E}}})) \wedge (\neg\, (\mu \overrightarrow{\mathtt{x}_{\mathscr{E}}} \in John)\, M(\overrightarrow{\mathtt{x}_{\mathscr{E}}})) \\
&= \min(0.5, 0.5) \\
&= 0.5
\end{aligned}$$

However, if we simply move μ above, we get the right conclusion.

$$\begin{aligned}
[\![\text{John is a musician and is not a musician}]\!] &= (\mu \overrightarrow{\mathtt{x}_{\mathscr{E}}} \in \mathit{John}) M(\overrightarrow{\mathtt{x}_{\mathscr{E}}}) \wedge \neg M(\overrightarrow{\mathtt{x}_{\mathscr{E}}}) \\
&= \frac{1}{n} \sum_{\overrightarrow{\mathtt{x}_{\mathscr{E}}} \in \mathsf{Ex}(\mathit{John}, n)} \min\left(M(\overrightarrow{\mathtt{x}_{\mathscr{E}}})), 1 - M(\overrightarrow{\mathtt{x}_{\mathscr{E}}})\right) \\
&= 0
\end{aligned}$$

For any concept C, this is true because $C(\overrightarrow{\mathtt{x}_{\mathscr{E}}})$ can only be 0 or 1 and $\min(0, 1 - 0) = \min(1, 1 - 1) = 0$. So, the core criticism against fuzzy logic is deftly avoided by simply assuming that the individual points are sampled at the matrix level rather than at each predicate.

Alternative definitions for conjunction and disjunctions

While the minimum and maximum are simple ways to define the connectives, there are in-fact an infinite number of possible functions we could use. Conjunction and disjunction can be represented in fuzzy logic more generally using functions called t-norms for conjunction or t-conorms for disjunction (Hájek 1998). A t-norm is a function, $T : [0,1] \times [0,1] \to [0,1]$, which satisfies the following conditions:

i. **Commutativity:** $T(x,y) = T(y,x)$

ii. **Associativity:** $T(x, T(y,z)) = T(T(x,y), z)$

iii. **Monotonicity:** $y \leq z \to T(x,y) \leq T(x,z)$

iv. **Identity element:** $T(x,1) = x$

The minimum satisfies these properties and is alternatively known as the Gödel t-norm after a logic he developed (Gödel 1932). Another example could be $T(x,y) = xy$, the product t-norm.

A t-conorm has all of the same conditions as a t-norm except for condition iv. Rather than $T(x,1) = x$, a t-conorm has $T(x,0) = x$. Likewise, the maximum is a t-conorm and is analogously called the Gödel t-conorm.

Finally, there is an even more general variant, known as a *uninorm* (Yager and Rybalov 1996). This allows for the identity element iv, to be any value $e \in [0,1]$. Therefore, the t-norm and t-conorm are just special cases where $e = 1$ and $e = 0$ respectively.

An example uninorm could be R_* where $e = 0.7$.

$$R_*(x,y) = \begin{cases} \max x, y & \text{if } x > 0.7 \text{ and } y > 0.7 \\ \min x, y & \text{else} \end{cases} \tag{2.13}$$

Essentially, this is the highest possible value when all elements are above 0.7. Otherwise, we choose the lowest possible value if any are below 0.7. So, above 0.7 we get behaviour like disjunction, and below like conjunction.

There are many advantages to this more general understanding of conjunction and disjunction in fuzzy logic. In particular, for linguistics, it allows for *partial content*. This is the idea that a proposition's truth value depends on its constituent parts in a non-classical way. For example $A \wedge B$ might *seem* more true than B alone if A is true and B is false. Contrast for instance, the following false sentences:

(14) The capital of Canada is Toronto.

(15) Canada is a large country located in North America which was founded in 1867, with a population of 38 million whose capital is Toronto.

Despite both (14) and (15) being false by classical logic, it is seems as though (15) is less false than (14). A different t-norm or uninorm might help capture this fact.

Chapter 3

The core fragment

3.1 Syntax

A conceptual language, $\mathscr{L}$ has the signature $\sigma = \langle \mathscr{P}, \mathscr{F}, \mathscr{K}, \mathscr{I}, \mathscr{C} \rangle$. The sets $\mathscr{F}$ and $\mathscr{P}$ are the usual function and predicate symbols, respectively.

Any category concept is a 1-place predicate in $\mathscr{P}$ which is also in $\mathscr{K}$. Any individual concept is a 1-place predicate in $\mathscr{P}$ which is also in $\mathscr{I}$. Any concept whether categorical or individual is in $\mathscr{C}$.

3.1.1 Alphabet

The language has the following alphabet, $\mathscr{A}$:

- All elements of $\mathscr{F} \cup \mathscr{P}$.
- $\neg, \wedge, \vee, \forall, \exists, \mu, \restriction, \approx, S, \ominus$
- $(,), \in$
- A countably infinite number of variables: $\overrightarrow{\mathtt{x}_{\mathscr{E}}}, \overrightarrow{\mathtt{y}_{\mathtt{R}}}, \ldots$

Let $\mathscr{A}^*$ be the set of all finite strings over $\mathscr{A}$.

3.1.2 Terms

Let $\mathscr{T} \subseteq \mathscr{A}^*$ be the set of well-formed terms for σ which satisfy the following conditions.

- If $\overrightarrow{\mathtt{x}_{\mathscr{E}}}$ is a variable, then $\overrightarrow{\mathtt{x}_{\mathscr{E}}} \in \mathscr{T}$.[1]

[1] Here the subscript reminds us that the variable is in the $\mathscr{E}$-space. We might require different spaces for some operations so we put a subscript to keep track of where a point is.

- If $c_{\mathscr{E}}$ is a constant symbol. then $c_{\mathscr{E}} \in \mathscr{T}$.

- If $f \in \mathscr{F}$ is a n-ary function symbol where $n \geq 1$ and $t_1, \ldots, t_n \in \mathscr{T}$, then $f(t_1, \ldots, t_n) \in \mathscr{T}$.

3.1.3 Formulae

Let $\mathscr{W} \subseteq \mathscr{A}^*$ be the set of well-formed formula for σ.

- If $P \in \mathscr{P}$ is a n-ary predicate symbol where $n \geq 1$ and $t_1, \ldots, t_n \in \mathscr{T}$, then $P(t_1, \ldots, t_n) \in \mathscr{W}$.

- If $C \in \mathscr{C}$ and $t \in \mathscr{T}$ then $C(t) \in \mathscr{W}$.

- If $\varphi \in \mathscr{W}$, then $(\neg \varphi) \in \mathscr{W}$.

- If $\varphi, \psi \in \mathscr{W}$, then $(\varphi \wedge \psi) \in \mathscr{W}$, $(\varphi \vee \psi) \in \mathscr{W}$, and $S(\varphi, \psi) \in \mathscr{W}$.

- If $\psi \in \mathscr{W}$, $C \in \mathscr{C}$ and $\overrightarrow{\mathrm{x}_{\mathscr{E}}}$ is a variable, then $((\forall \overrightarrow{\mathrm{x}_{\mathscr{E}}} \in C)\ \psi) \in \mathscr{W}$, $((\exists \overrightarrow{\mathrm{x}_{\mathscr{E}}} \in C)\ \psi) \in \mathscr{W}$, $((\mu \overrightarrow{\mathrm{x}_{\mathscr{E}}} \in C)\ \psi) \in \mathscr{W}$.

- If $t_1, t_2 \in \mathscr{T}$ and $C \in \mathscr{C}$ then $(t_1 \underset{C}{\approx} t_2) \in \mathscr{W}$.

- If $A, B \in \mathscr{C}$ and $t \in \mathscr{T}$ then $(A \upharpoonright B)\,(t) \in \mathscr{W}$.

3.1.4 Concepts

Here we define $\mathscr{C}$ inductively.

- If $C \in \mathscr{K}$ then $C \in \mathscr{C}$.

- If $C \in \mathscr{I}$ then $C \in \mathscr{C}$.

- Let $C \in \mathscr{C}$ and $\varphi \in \mathscr{W}$. If φ has exactly one free variable, $\overrightarrow{\mathrm{x}_{\mathscr{E}}}$, then $\ominus(C, \varphi) \in \mathscr{C}$.

3.2 Semantics

We can define a model for $\mathscr{L}$ with the following 4-tuple, $\langle \sigma, \mathscr{S}, D, I \rangle$. We have our signature, σ as well as the set of spaces, $\mathscr{S}$, our domain D, and our interpretation function, I.

Until now, we have only been working within a single conceptual space, the $\mathscr{E}$-space. To handle adjectives and the computational model, we will need *multiple* conceptual spaces. This will be used to handle scales like “height”, for example. $\mathscr{S}$ is a set of conceptual spaces, for example $\mathscr{E} \in \mathscr{S}$.

Since we have multiple spaces, we need to make sure that if we have a point, we know which spaces it is in. To accomplish this, we always index our points to a specific space, $s \in \mathscr{S}$. This lets us know which predicates and functions can apply to it. So, the domain of discourse, D, is a set of ordered pairs consisting of a conceptual space $s \in \mathscr{S}$, and a vector in $\mathbb{R}^n$, where n is an arbitrary integer. We call these pairs “situated individuals” since they are situated in a specific conceptual-space.

$$D = \{\langle s, \vec{x} \rangle \mid s \in \mathscr{S}, \vec{x} \in \mathbb{R}^n\}$$

To keep track of which spaces that our variables or functions operate in, we notate the conceptual space with a subscript. For example, $\overrightarrow{\mathrm{x}_{\mathscr{E}}}$ is a variable that is restricted to points in the $\mathscr{E}$-space.

3.2.1 Propositions

Ultimately, our system will be *dynamic* and the rules for a monotonic update will be discussed in Chapter 6. Nevertheless, the static representation of a proposition is the set of models which satisfy the proposition. Since the logic is fuzzy, we restrict this meaning to be only the models which satisfy a proposition with a truth value of 1.

3.2.2 Interpretation of connectives

For the basic connectives, ¬, ∧ and ∨, we use the standard fuzzy logic interpretations (Zadeh 1965).

$$I(\neg\varphi) = 1 - I(\varphi) \tag{3.1}$$

$$I(\varphi \wedge \psi) = \min(I(\varphi), I(\psi)) \tag{3.2}$$

$$I(\varphi \vee \psi) = \max(I(\varphi), I(\psi)) \tag{3.3}$$

See Section 2.6 for a more involved discussion of alternatives.

3.2.3 Predicates in $\mathscr{E}$-space

Each space will have different interpretations for their situated individuals reflecting the different computational necessities for each space. I will begin with the primary space, $\mathscr{E}$, before discussing adjective spaces in Chapter 5. Here, we will discuss primarily predicates that are in $\mathscr{C}$ and are concept-triples.

3.2.4 Structure of $\mathscr{E}$-predicates

Each concept, $C \in \mathscr{C}$ is fundamentally a triple, which lives in $\mathscr{E}$.

$$I(C) = \langle M_P, P_P, g_P \rangle \tag{3.4}$$

For any concept $C \in \mathscr{C}$, we can define a distance, d_C between two points on the manifold using g_P and M_P (c.f. Section 2.3.4).

Interpretation

The interpretation of $C \in \mathscr{C}$ as a predicate checks if $\overrightarrow{\mathrm{x}_{\mathscr{E}}}$ lies on the manifold of C or not.

$$I(C(\overrightarrow{\mathrm{x}_{\mathscr{E}}})) = \begin{cases} 1 & \text{if } \overrightarrow{\mathrm{x}_{\mathscr{E}}} \in M_C \\ 0 & \text{if } \overrightarrow{\mathrm{x}_{\mathscr{E}}} \notin M_C \end{cases} \tag{3.5}$$

3.2.5 Quantifiers

Let C be any concept triple, and $\varphi(\overrightarrow{\mathrm{x}_{\mathscr{E}}})$ be an open proposition. To sample points from a concept $C \in \mathscr{C}$, we use $\mathsf{Ex}\,(C,n)$ which is a set of n points independently sampled from P_C.

$$I\left((\mu \overrightarrow{\mathrm{x}_{\mathscr{E}}} \in C)\,\varphi\right) = \frac{1}{n} \sum_{\overrightarrow{\mathrm{x}_{\mathscr{E}}} \in \mathsf{Ex}(C,n)} \varphi(\overrightarrow{\mathrm{x}_{\mathscr{E}}}) \tag{3.6}$$

$$I\left((\exists \overrightarrow{\mathrm{x}_{\mathscr{E}}} \in C)\,\varphi\right) = \max_{\overrightarrow{\mathrm{x}_{\mathscr{E}}} \in \mathsf{Ex}(C,n)} \varphi(\overrightarrow{\mathrm{x}_{\mathscr{E}}}) \tag{3.7}$$

$$I\left((\forall \overrightarrow{\mathrm{x}_{\mathscr{E}}} \in C)\,\varphi\right) = \min_{\overrightarrow{\mathrm{x}_{\mathscr{E}}} \in \mathsf{Ex}(C,n)} \varphi(\overrightarrow{\mathrm{x}_{\mathscr{E}}}) \tag{3.8}$$

3.2.6 Individuals

For any individual concept $C \in \mathscr{I}$, and an open proposition, $\varphi(\overrightarrow{\mathrm{x}_{\mathscr{E}}})$, we predicate it by simply using μ.

$$I(\varphi(C)) = I\left((\mu \overrightarrow{\mathrm{x}_{\mathscr{E}}} \in C)\,\varphi(\overrightarrow{\mathrm{x}_{\mathscr{E}}})\right)$$

3.2.7 Concept coercion

Sometimes there isn't a lexical concept-triple, $\langle M_P, P_p, g_P \rangle$ (e.g. when we restrict a noun with a relative clause or adjective). In those cases, we can coerce a concept if we want to look at distances or sample from the restricted concept. This is necessary if we want to be able to quantify over complex concepts.

The operator $\ominus$ allows us to take arbitrary open propositions, $\varphi(\overrightarrow{\mathtt{x}_{\mathscr{E}}})$, and coerce them into submanifolds of a concept, C, so that they have a concept-triple. An open proposition is just any proposition φ where there is one variable, $\overrightarrow{\mathtt{x}_{\mathscr{E}}}$ which is left unbound.

$$I(\ominus(C,\varphi)) = \left\langle M_{\ominus(C,\varphi)}, P_{\ominus(C,\varphi)}, g_{\ominus(C,\varphi)} \right\rangle \tag{3.9}$$

$$M_{\ominus(C,\varphi)} = \{\overrightarrow{\mathtt{x}_{\mathscr{E}}} \in M_C | I(\varphi(\overrightarrow{\mathtt{x}_{\mathscr{E}}})) = 1\} \tag{3.10}$$

$$P_{\ominus(C,\varphi)} = \frac{P_C(\overrightarrow{\mathtt{x}_{\mathscr{E}}})}{\int_{M_{\ominus(C,\varphi)}} P_C(\overrightarrow{\mathtt{x}_{\mathscr{E}}}) d\overrightarrow{\mathtt{x}_{\mathscr{E}}}} \tag{3.11}$$

$$g_{\ominus(C,\varphi)} = g_C \tag{3.12}$$

The manifold of $\ominus(C,\varphi)$ is all points, $\overrightarrow{\mathtt{x}_{\mathscr{E}}}$, on the manifold of C for which $\varphi(\overrightarrow{\mathtt{x}_{\mathscr{E}}})$ is true. The probability distribution of $\ominus(C,\varphi)$ is P_C renormalised over the new manifold, and its metric tensor is g_C. In reality, the probability distribution and metric tensor would be affected by the constituents of φ, but this works as an approximation. For instance, when presented with the complex concept of "Harvard-educated carpenters", people assume that such an individual would be idealistic (Hampton 1987); it's not clear that domain restriction alone could handle this.

$\ominus(C,\varphi)$ is a concept triple in $\mathscr{C}$ just like C and so it can appear anywhere C can appear.

3.3 Examples

We can now look at some concrete examples. Let's take a simple sentence like "John is a man".

$$\begin{aligned}
[\![\text{John is a man}]\!] &= [\![\text{John}]\!]\left([\![\text{is a man}]\!]\right) \\
&= \left(\lambda P.\left(\mu\, \overrightarrow{\mathtt{j}}_{\mathscr{E}} \in John\right) P(\overrightarrow{\mathtt{j}_{\mathscr{E}}})\right)\left([\![\text{is a man}]\!]\right) \\
&= \left(\mu\, \overrightarrow{\mathtt{j}}_{\mathscr{E}} \in John\right) [\![\text{is a man}]\!]\left(\overrightarrow{\mathtt{j}_{\mathscr{E}}}\right) \\
&= \left(\mu\, \overrightarrow{\mathtt{j}}_{\mathscr{E}} \in John\right) Man(\overrightarrow{\mathtt{j}_{\mathscr{E}}}) \\
&= \frac{1}{n} \sum_{\overrightarrow{\mathtt{j}_{\mathscr{E}}} \in \mathsf{Ex}(John,n)} \begin{cases} 1 & \text{if } \overrightarrow{\mathtt{j}_{\mathscr{E}}} \in M_{Man} \\ 0 & \text{if } \overrightarrow{\mathtt{j}_{\mathscr{E}}} \notin M_{Man} \end{cases}
\end{aligned}$$

How about “John is a man and not a woman”:

$$\begin{aligned}
\llbracket\text{John is a man and not a woman}\rrbracket &= \llbracket\text{John}\rrbracket\left(\llbracket\text{is a man and not a woman}\rrbracket\right)\\
&= \left(\mu\,\overrightarrow{\mathrm{j}}_{\mathscr{E}} \in John\right)\llbracket\text{is a man and not a woman}\rrbracket\left(\overrightarrow{\mathrm{j}_{\mathscr{E}}}\right)\\
&= \left(\mu\,\overrightarrow{\mathrm{j}}_{\mathscr{E}} \in John\right)Man(\overrightarrow{\mathrm{j}_{\mathscr{E}}}) \wedge \neg Woman(\overrightarrow{\mathrm{j}_{\mathscr{E}}})\\
&= \frac{1}{n}\sum_{\overrightarrow{\mathrm{j}_{\mathscr{E}}} \in \mathsf{Ex}(John,n)} \min\left(Man(\overrightarrow{\mathrm{j}_{\mathscr{E}}}), 1 - Woman(\overrightarrow{\mathrm{j}_{\mathscr{E}}})\right)
\end{aligned}$$

What about if we have a more complex concept?

$$\llbracket\text{All tall men who run are healthy.}\rrbracket = \left(\forall\overrightarrow{\mathrm{x}_{\mathscr{E}}} \in \ominus\left(Men, \llbracket\text{Tall}\rrbracket\left(\overrightarrow{\mathrm{x}_{\mathscr{E}}}\right) \wedge \llbracket\text{Run}\rrbracket\left(\overrightarrow{\mathrm{x}_{\mathscr{E}}}\right)\right)\right)\llbracket\text{Healthy}\rrbracket\left(\overrightarrow{\mathrm{x}_{\mathscr{E}}}\right)$$

3.4 The Semantics of Similarity

3.4.1 Step or sigmoid function

S remaps a value from $[0,1] \to [0,1]$. Specifically, S is a step function:

$$I(S(\varphi,\psi)) = \begin{cases} 1 & \text{if } I(\varphi) \geq I(\psi) \\ 0 & \text{if } I(\varphi) < I(\psi) \end{cases} \tag{3.13}$$

The step function remaps propositions with truth-values between 0 and 1 to just 1 or 0. This is useful if we ever want to take a predicate with a truth value that isn’t 1 or 0 and make it binary. For instance, if we wanted to turn a similarity statement into a manifold, we would need a cut-off point to define the boundaries of the manifold.

Often, we would pass specific truth-values for ψ, to do things like $S(\varphi, 0.5)$ where ‘0.5’ would stand for a 0-place predicate which has a truth value of 0.5. Alternatively, it could be used to check whether a proposition is as true as another proposition. In a computational setting such as Section 3.6, we would use a *sigmoid* function instead which is analogous to the step function but is continuous and differentiable throughout.

3.4.2 Similarity

To evaluate similarity, we use $\underset{C}{\approx}$ where C is a concept-triple and $\overrightarrow{\mathrm{x}_{\mathscr{E}}}, \overrightarrow{\mathrm{y}_{\mathscr{E}}}$ are terms in the $\mathscr{E}$-space (c.f. Section 2.3.3).

$$\overrightarrow{\mathrm{x}_{\mathscr{E}}} \underset{C}{\approx} \overrightarrow{\mathrm{y}_{\mathscr{E}}} = e^{-d_C(\overrightarrow{\mathrm{x}_{\mathscr{E}}},\overrightarrow{\mathrm{y}_{\mathscr{E}}})} \tag{3.14}$$

3.4.3 Metaphorical $\mathscr{E}$-predicates

As discussed in Section 2.3.5, words can easily be used in a metaphorical sense. We get metaphor predicates by adding the special "metaphorical" connective $(A \upharpoonright B)$ where A and B are concept triples.

$$I((A \upharpoonright B)\,(\overrightarrow{\mathtt{x}_{\mathscr{E}}})) = S\left(\max_{\overrightarrow{\mathtt{y}_{\mathscr{E}}} \in M_A}\left(\overrightarrow{\mathtt{x}_{\mathscr{E}}} \underset{B}{\approx} \overrightarrow{\mathtt{y}_{\mathscr{E}}}\right), \alpha\right) \tag{3.15}$$

This finds how similar $\overrightarrow{\mathtt{x}_{\mathscr{E}}}$ is to the entire A manifold, using B for a distance metric. We take whichever point in A that $\overrightarrow{\mathtt{x}_{\mathscr{E}}}$ is most similar to, and use that to determine how true the metaphor is.

Finally, we wrap the entire thing in a step function to remap from similarities in $(0,1]$ to either 1 or 0. α is a free variable in $[0,1]$ which simply determines the cut-off point for the metaphor.

3.5 More examples

Let's say that John and Bob are similar musicians.

$$\begin{aligned}
\llbracket\text{John is like Bob}\rrbracket &= \llbracket\text{John}\rrbracket\,(\llbracket\text{is like}\rrbracket\,(\llbracket\text{Bob}\rrbracket))\\
&= \left(\lambda P.\left(\mu\overrightarrow{\mathtt{j}_{\mathscr{E}}} \in John\right) P(\overrightarrow{\mathtt{j}_{\mathscr{E}}})\right)\\
&\quad \left(\left(\lambda T.\lambda\overrightarrow{\mathtt{x}_{\mathscr{E}}}.T(\lambda\overrightarrow{\mathtt{y}_{\mathscr{E}}}.\overrightarrow{\mathtt{x}_{\mathscr{E}}} \underset{C}{\approx} \overrightarrow{\mathtt{y}_{\mathscr{E}}})\right)\left(\lambda P\left(\mu\overrightarrow{\mathtt{b}_{\mathscr{E}}} \in Bob\right) P(\overrightarrow{\mathtt{b}_{\mathscr{E}}})\right)\right)\\
&= \left(\lambda P.\left(\mu\overrightarrow{\mathtt{j}_{\mathscr{E}}} \in John\right) P(\overrightarrow{\mathtt{j}_{\mathscr{E}}})\right)\\
&\quad \left(\lambda\overrightarrow{\mathtt{x}_{\mathscr{E}}}.\left(\lambda P\left(\mu\overrightarrow{\mathtt{b}_{\mathscr{E}}} \in Bob\right) P(\overrightarrow{\mathtt{b}_{\mathscr{E}}})\right)(\lambda\overrightarrow{\mathtt{y}_{\mathscr{E}}}.\overrightarrow{\mathtt{x}_{\mathscr{E}}} \underset{C}{\approx} \overrightarrow{\mathtt{y}_{\mathscr{E}}})\right)\\
&= \left(\lambda P.\left(\mu\overrightarrow{\mathtt{j}_{\mathscr{E}}} \in John\right) P(\overrightarrow{\mathtt{j}_{\mathscr{E}}})\right)\\
&\quad \left(\lambda\overrightarrow{\mathtt{x}_{\mathscr{E}}}.\left(\mu\overrightarrow{\mathtt{b}_{\mathscr{E}}} \in Bob\right)\overrightarrow{\mathtt{x}_{\mathscr{E}}} \underset{C}{\approx} \overrightarrow{\mathtt{b}_{\mathscr{E}}}\right)\\
&= \left(\mu\overrightarrow{\mathtt{j}_{\mathscr{E}}} \in John\right)\left(\lambda\overrightarrow{\mathtt{x}_{\mathscr{E}}}.\left(\left(\mu\overrightarrow{\mathtt{b}_{\mathscr{E}}} \in Bob\right)(\overrightarrow{\mathtt{x}_{\mathscr{E}}} \underset{C}{\approx} \overrightarrow{\mathtt{b}_{\mathscr{E}}})\right)\right)(\overrightarrow{\mathtt{j}_{\mathscr{E}}})\\
&= \left(\mu\overrightarrow{\mathtt{j}_{\mathscr{E}}} \in John\right)\left(\left(\mu\overrightarrow{\mathtt{b}_{\mathscr{E}}} \in Bob\right)(\overrightarrow{\mathtt{j}_{\mathscr{E}}} \underset{C}{\approx} \overrightarrow{\mathtt{b}_{\mathscr{E}}})\right)\\
&= \frac{1}{n}\sum_{\overrightarrow{\mathtt{j}_{\mathscr{E}}} \in \mathsf{Ex}(John,n)} \frac{1}{n}\sum_{\overrightarrow{\mathtt{b}_{\mathscr{E}}} \in \mathsf{Ex}(Bob,n)} e^{-d_C(\overrightarrow{\mathtt{j}_{\mathscr{E}}},\overrightarrow{\mathtt{b}_{\mathscr{E}}})}
\end{aligned}$$

The free variable, C would be filled in by *Musicians* in this context.

What about if we want to refer to someone as a lion as a fighter (e.g. their strength or valour)? Here we use the metaphorical operator, $\upharpoonleft$.

$$\begin{aligned}
[\![\text{John is a lion}]\!] &= [\![\text{John}]\!]\,([\![\text{is a lion}]\!]) \\
&= \left(\lambda P.\left(\mu\,\overrightarrow{\mathtt{j}}_{\mathscr{E}} \in \mathit{John}\right) P(\overrightarrow{\mathtt{j}_{\mathscr{E}}})\right)([\![\text{is a lion}]\!]) \\
&= \left(\mu\,\overrightarrow{\mathtt{j}}_{\mathscr{E}} \in \mathit{John}\right) [\![\text{is a lion}]\!]\,(\overrightarrow{\mathtt{j}_{\mathscr{E}}}) \\
&= \left(\mu\,\overrightarrow{\mathtt{j}}_{\mathscr{E}} \in \mathit{John}\right) \left((\mathit{Lion} \upharpoonleft C)\,(\overrightarrow{\mathtt{j}_{\mathscr{E}}})\right) \\
&= \frac{1}{n} \sum_{\overrightarrow{\mathtt{j}_{\mathscr{E}}} \in \mathsf{Ex}(\mathit{John},n)} S\left(\max_{\overrightarrow{\mathtt{x}_{\mathscr{E}}} \in \mathit{Lion}} e^{-d_C(\overrightarrow{\mathtt{j}_{\mathscr{E}}},\overrightarrow{\mathtt{x}_{\mathscr{E}}})}, \alpha\right)
\end{aligned}$$

3.5.1 Individual and category coercion

Sentence (16) has three readings.

(16) John is like a lion.

i) There is a specific lion that John is similar to.

ii) There exists some lion that John is similar to.

iii) John is similar to lions in general.

To handle iii), we can treat the "lion" concept as an individual. While "lion" is not in the set of individual concepts, $\mathscr{I}$, it is the same kind of object (a concept triple), so the coercion is trivial.

$$\begin{aligned}
[\![\text{John is like a lion}]\!] &= \left(\mu\overrightarrow{\mathtt{j}_{\mathscr{E}}} \in \mathit{John}\right) \left(\left(\mu\overrightarrow{\mathtt{l}_{\mathscr{E}}} \in \mathit{Lion}\right) (\overrightarrow{\mathtt{j}_{\mathscr{E}}} \underset{C}{\approx} \overrightarrow{\mathtt{l}_{\mathscr{E}}})\right) \\
&= \frac{1}{n} \sum_{\overrightarrow{\mathtt{j}_{\mathscr{E}}} \in \mathsf{Ex}(\mathit{John},n)} \frac{1}{n} \sum_{\overrightarrow{\mathtt{l}_{\mathscr{E}}} \in \mathsf{Ex}(\mathit{Lion},n)} e^{-d_C(\overrightarrow{\mathtt{j}_{\mathscr{E}}},\overrightarrow{\mathtt{l}_{\mathscr{E}}})}
\end{aligned}$$

In fact, the reading iii) of (16) is an indefinite singular generic. This kind of category to individual coercion will be quite useful in Chapter 4.

We can also coerce individual concepts to concept category concepts. Someone who is very smart can be referred to as "an Einstein"; this is simply our applying our metaphorical operator to the Einstein individual!

$$[\![\text{John is an Einstein}]\!] = \left(\mu\overrightarrow{\mathtt{j}_{\mathscr{E}}} \in \mathit{John}\right) (\mathit{Einstein} \upharpoonleft \mathit{Thinker}) \left(\overrightarrow{\mathtt{j}_{\mathscr{E}}}\right)$$

The fact that Einstein is a brilliant thinker produces the intended meaning. Of course, such compositional metaphors might end up becoming fossilised as idioms or as lexical items.

3.6 A computational model

So far, I have described the system in entirely formal terms. However, while infinite-dimensional manifolds or pathological functions are of interest to the mathematician, the cognitive scientist ought to be interested in algorithms that can be explicitly calculated. Indeed, a major problem with formal approaches to cognitive science is that many are computationally intractable. While there have been attempts to side-step the computational intractability as irrelevant, these just-so stories are far from satisfactory. (I. v. Rooij et al. 2018).

For this system in particular, there is no immediately clear way to represent a manifold since they are an incredibly diverse set of mathematical objects. Furthermore, while Chapter 6 will address dynamic reasoning, it does so assuming that the manifolds in question already existed and can then simply be manipulated. The goal of this section is to outline a computational model which can directly implement the different symbols of the language described here in a computationally feasible way. Furthermore, it does not operate over manifolds which have been defined already but rather implements them in a way such that they can be *learnt*. This serves as a proof of concept for learning with these representations.

Concretely, the model learns the appropriate representations to make an arbitrary set of propositions true. In this regard, it is a neuro-symbolic model (Garcez and Lamb 2020) which are models which integrate formal symbolic representations with neural networks. This model will highlight possible different connections between parts of the model and certain constraints on concepts that could exist for humans as well.

3.6.1 Representing the concept manifold

We first greatly reduce the set of possible manifolds to those which are diffeomorphic to the unit ball of $\mathbb{R}^n$ (the unit ball is $\{\vec{x} \in \mathbb{R}^n | \; d(\vec{x}, 0) < 1\}$) where n is the number of dimensions of the $\mathscr{E}$-space. If we have two manifolds, M and N, then a function $f : M \to N$ is a diffeomorphism if it is bijective and both f and f^{-1} are differentiable.

In other words, while our manifolds can have all sorts of possible shapes, there must always be a way to smoothly transform it to the unit ball. This assumption immediately enforces both of Gardenförs's convexity and connectedness constraint, while allowing our representations to take on radically different shapes. However, no concept-manifold could be disconnected or have a hole because each manifold is diffeomorphic to the unit ball.

For a concept C, we define a function $f_{\mathsf{C}} : \mathscr{E} \to \mathsf{C}$, where C is a conceptual space *just* for the concept C. This takes a point in $\mathscr{E}$ and remaps it to a conceptual space, C. In C, any point within 1 unit from the origin is considered a member of

C, anything further than that is considered to not be a member of C. This function is invertible, so we can also map points from C to $\mathcal{E}$ with f_{C}^{-1}.

Invertible neural networks

To represent this function, we use an *invertible neural network*. Invertible neural networks are a class of neural networks which are defined to be always bijective and map from $\mathbb{R}^n$ to $\mathbb{R}^n$ (Dinh, Krueger, and Bengio 2015; Kingma and Dhariwal 2018; Behrmann et al. 2019). By initialising f_{C} as an invertible neural network, we can learn a unique mapping for each concept C from $\mathcal{E}$ to C. Crucially this mapping will be continuous and differentiable since it is a neural network. Then, the manifold of a concept in $\mathcal{E}$ is $\{x \in \mathcal{E} | d(f_{\mathsf{C}}(x), \vec{0}) < 1\}$.

$$M_C = \{x \in \mathcal{E} | d(\vec{0}, f_{\mathsf{C}}(x)) < 1\} \tag{3.16}$$

3.6.2 The metric

To define the metric distance, we calculate distances inside the unit-ball. Concretely, the metric tensor is the same object for both the manifold embedded in $\mathcal{E}$-space or inside the unit-ball, but it has a different expression in coordinates. For example, distances on the Earth don't depend on whether you use the Mercator projection, Robison projection or even Buckminster Fuller's rather strange Dymaxion map. However, it is much easier to measure distances on some maps than others.

So, while the distance along the manifold might have a closed-form expression in C, it might not inside $\mathcal{E}$. Since the length of a curve is invariant under a change of coordinates (Ricci and Levi-Civita 1900), we can measure distances in C in constant time if we have such a closed-form solution, to determine lengths in $\mathcal{E}$. As such, we could measure distance inside the unit-ball with Euclidean distances or with a hyperbolic distance (modeled with a Poincaré-ball). For example, assuming we measure with the Euclidean distance inside C:

$$d_C(\overrightarrow{\mathrm{x}_{\mathcal{E}}}, \overrightarrow{\mathrm{y}_{\mathcal{E}}}) = \inf(\{\text{continuous curves from } \overrightarrow{\mathrm{x}_{\mathcal{E}}} \text{ to } \overrightarrow{\mathrm{y}_{\mathcal{E}}} \text{ on } M_C\}) = \sqrt{\sum_i^n (f_{\mathsf{C}}(y_{\mathcal{E}})_i - f_{\mathsf{C}}(\overrightarrow{\mathrm{x}_{\mathcal{E}}})_i)^2} \tag{3.17}$$

3.6.3 The probability distribution

To define the probability distribution over P_C, we define a probability distribution over the unit ball. Then, for any point we sample on the unit-ball, we can get its position in the $\mathcal{E}$-space with f_{C}^{-1}.

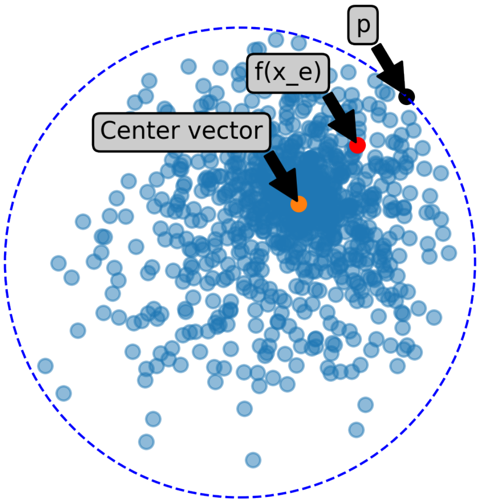


Figure 3.1: Internal structure of a conceptual manifold. Points are sampled from the distribution on the manifold. An exemplar, $\vec{x_C}$, is sampled by choosing a direction and then sampling between the centre and p according to a $B(1,b)$ where b is a learnt parameter between 1 and 3 and B is the beta distribution

An example of a family of such distributions could be the following:

There is a learnt vector (constrained to the unit ball) which represents the "centre" of the concept. A direction, $\vec{d}$ is uniformly sampled from the unit sphere. We find the line segment from the centre and the boundary of the ball in direction $\vec{d}$. We then sample along that line segment according to $l \sim B(1,b)$ where b is a learnt parameter between 1 and 3 and B is the beta distribution. This gives us a family of probability distributions for each manifold. Figure 3.1 shows this structure.

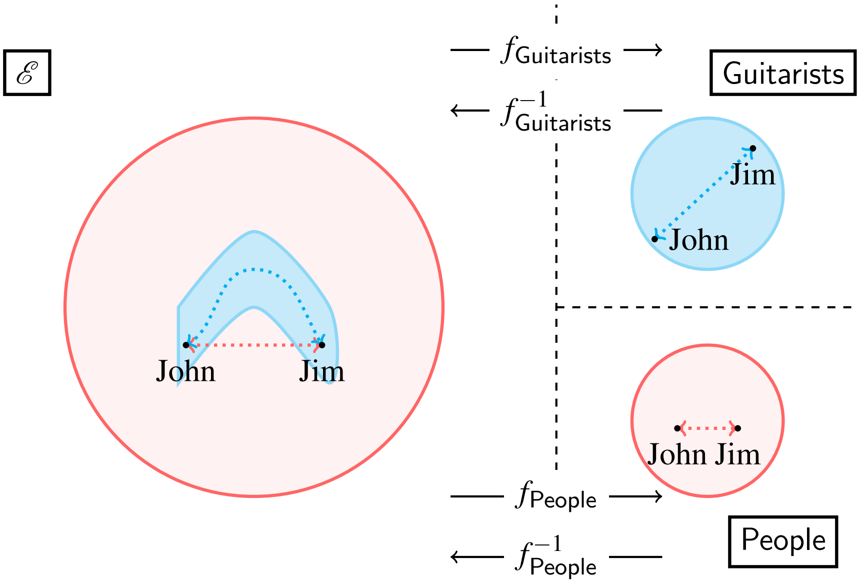
This will force the distribution to be unimodal, but in practice we should easily imagine multimodal distributions for concepts.

3.6.4 The loss function

$$Loss(\varphi) = -\log\varphi \tag{3.18}$$

To learn to model a given proposition with gradient-descent, we simply take the negative log of the truth value of that proposition. If the loss is 0, then the proposition has a truth value of one. Otherwise it gets increasingly large as we approach a truth value of zero.

All connectives are simply used as defined in this chapter. Since predication



$\mathscr{E}$
$f_{\mathsf{Guitarists}}$
$f^{-1}_{\mathsf{Guitarists}}$
Guitarists
Jim
John
John
Jim
John Jim
f_{People}
f^{-1}_{People}
People

Figure 3.2: Replication of Figure 2.5 to show how the internal distance calculation works. We take points in $\mathscr{E}$ to either People or Guitarists using f_{People} or $f_{\mathsf{Guitarists}}$. Rather than needing to do the complicated curve for guitarists in the $\mathscr{E}$, we have a very simple straight line in the unit-ball which directly corresponds to the same curve.

is not differentiable, we need to switch it out with a smooth approximation.

$$P(\overrightarrow{\mathrm{x}_{\mathscr{E}}}) = \begin{cases} 1 & \text{if } d(f_{\mathsf{P}}(\overrightarrow{\mathrm{x}_{\mathscr{E}}}),\vec{0}) < 1 \\ 0 & \text{if } d(f_{\mathsf{P}}(\overrightarrow{\mathrm{x}_{\mathscr{E}}}),\vec{0}) \geq 1 \end{cases} \tag{3.19}$$

The solution is to simply smoothly approximate Equation 3.19 using a sigmoid function.

$$P(\overrightarrow{\mathrm{x}_{\mathscr{E}}}) = \frac{1}{1 + e^{\alpha(1 - d(f_{\mathsf{P}}(\overrightarrow{\mathrm{x}_{\mathscr{E}}}),\vec{0}))}} \tag{3.20}$$

This will have a value above 0.5 for anything within P and a value below 0.5 for anything outside of P. If we increase α (a chosen hyperparameter rather than a learnt value), we can make the transition much sharper, so that anything even a little bit inside P is close to 1, and anything outside is close to 0.

Alternatively, we can integrate our probability distribution so that the maximum of $P(\overrightarrow{\mathrm{x}_{\mathscr{E}}})$ is at the centre of P. Let $\overrightarrow{\mathrm{c}_{\mathsf{C}}}$ be the centre of our probability distribution. Let $\overrightarrow{\mathrm{z}_{\mathsf{C}}}$ be the closest point on the unit sphere which lies on the line between the centre and $f_{\mathsf{P}}(\overrightarrow{\mathrm{x}_{\mathscr{E}}})$. Note that if $\overrightarrow{\mathrm{c}_{\mathsf{C}}} = 0$, then Equation 3.21 reduces to Equation 3.20.

$$P(\overrightarrow{\mathtt{x}_{\mathscr{E}}}) = \frac{1}{1 + e^{\alpha(d(\overrightarrow{\mathtt{z}_{\mathsf{C}}},\overrightarrow{\mathtt{c}_{\mathsf{C}}}) - d(f_{\mathsf{P}}(\overrightarrow{\mathtt{x}_{\mathscr{E}}}),\overrightarrow{\mathtt{c}_{\mathsf{C}}}))}} \tag{3.21}$$

This connects our probability distribution to our predicate classification in a structured way; things which are easy to classify (have low loss) are also those things which are sampled the most. If we had a different class of probability distribution, we would need a different loss function, however. This system is actually implemented in Python with Pytorch in this GitHub repo[2] which contains some example models.

[2] https://github.com/MichaelGoodale/conceptual-manifolds

Chapter 4

Generics

Generics represent basic generalisations about categories. There have been many vastly different approaches to generics. Despite this large literature, there is little in the way of consensus about the interpretation of generics. Analyses include a special kind of quantifier (Lewis 1975; Krifka, Pelletier, et al. 1995), predication over kinds (Carlson 1977; Liebesman 2011), cognitive "defaults" (Leslie 2007), possible world semantics (Asher and Morreau 1995), and probabilistic approaches (Cohen 1996; R. v. Rooij and Schulz 2020; Kochari, R. v. Rooij, and Schulz 2020). Many of these theories add different types of generics with separate meanings.

Furthermore, languages sometimes have different kinds of generic sentences. For example, each sentence of (17) has a reading which means something like (17a).

(17) a. Lions have manes.

b. The lion has a mane.

c. A lion has a mane.

This chapter will give the formal foundations for a theory of generics. It will address indefinite singular generics, which have received considerably less attention in the literature than bare plurals like (17a) (Leslie and Lerner 2016). Unlike bare plurals or definite generics, indefinite singular generics cannot take direct-kind predicates which are true for the kind, but not any individual (e.g. "extinct"). For example, all of (18) are infelicitous or have very different meanings than the bare-plural equivalent.

(18) a. ? A dodo bird is extinct.

b. ? A mosquito is widespread.

c. ? A car causes pollution all around the world.

They also cannot take collective predicates.

(19) a. ? A person gathers every year to celebrate New Year's Eve.

b. ? A cow emits forty percent of global methane emissions.

A common naïve idea for generics is that they reflect things which are true for a majority of the members of a concept.

(20) A dog has four legs.

Nearly all dogs have four legs, and so (20) is true. This majority explanation doesn't stand up to the slightest bit of scrutiny however. For example, most ducks don't lay eggs, considering male and juvenile ducks, and yet (21) is true.

(21) A duck lays eggs.

Worse still, there are more female ducks than egg-laying ducks but (22) sounds bizarre.

(22) ? A duck is female.

A common view is that indefinite generics describe either necessary or essential properties of a kind (Lawler 1974). This is often analysed as generics representing "rules or regulations" or partial definitions (Cohen 2001). So, (23) is rejected since being right-handed is not a necessary condition for being a man.

(23) ? A man is right-handed.

However, this doesn't work either since laying eggs is by no means a necessary condition for being a duck. Furthermore, there are some dogs which have only three legs! Leslie 2007 argues that generics (bare-plural or indefinite) represent default cognitive generalisations about a concept, although she doesn't offer a specific formal analysis of generics. This in part because she thinks the generalisations are so cognitively primitive that they could not be accounted for by a quantificational analysis.

I agree with Leslie that generics reflect default generalisation but I do not think this weds us to a non-quantificational analysis. Rather, we can account for these kinds of generics, using the tools developed in Chapters 2 and 3.

4.1 Indefinite singular generics are coerced individuals

Concretely, the idea is simple: indefinite singular generics are category concepts coerced to individual concepts. Since individual concepts and category concepts have identical internal structure, this type coercion is trivial. Finally, μ is what allows us simple predication over individuals, *and yet* it behaves just like a quantifier since it allows us to quantify over points from a concept. Note that μ quantifies over *points*, which are not immediately linguistically accessible (we cannot verbalise a point, only concepts which have sets of points). Conversely, quantifiers in natural language like "every" quantify over individuals which are linguistically accessible. Furthermore, μ is phonologically null in both simple predication and thus in generic statements.

$$[\![\text{A lion has a mane}]\!] = (\mu \overrightarrow{\mathrm{x}_{\mathscr{E}}} \in \mathit{Lion})\mathit{HasMane}(\overrightarrow{\mathrm{x}_{\mathscr{E}}}) \tag{4.1}$$

As Leslie argues, generics are true in virtue of our default assumptions. In this system, these default assumptions come from the regions of the concept with the highest probability of being sampled. The exemplar sampling automatically reduces us to the most typical members of a concept, obviating the need for any special new machinery for generics.

4.2 Sampling and generics

Generic sentences reflect the structure of the concepts that involve them. In particular, they reflect those claims we are willing to accept about a given concept that are deeply integrated into our understanding of the concept that goes beyond just its extension. As such, they can have a incredibly strong normative force. Even though we know that not all birds fly, it does seem that *a bird which does fly* is a better example of a bird than one that does not. *A duck must lay eggs* in order for the species to survive, whereas a duck being male or female has no bearing on its ducky-ness. This normative force, of course, can have profound social consequences, indeed part of the reason (23) is inadmissible is that it seems to imply that people *ought* to be right-handed and that left-handed people would otherwise be aberrant. These generic sentences show the descriptive properties that are integral to a concept. The probability distribution of a concept will maximise the probability mass in order to make these kinds of important, teleological properties of concepts true, even if they do not define the extension of that concept. Of course, if any property is true for all members of a concept, then that generic will of course be licenced as well.

(24) ? A book is a paperback.

There is no social reason to deny the fact that most books are paperbacks, yet (24) is false since we do not sample paperbacks at a higher rate than hardcovers when sampling points from the book concept. This is because P_{book} is *not* a probability distribution that models beliefs about the frequency of encountering different types. Rather, it models the likelihood of sampling an exemplar. The probability distribution is the source of the "default" generalisations that Leslie refers to; this is also why people sometimes endorse generics prefaced with "all" (c.f. Section 2.4.1).

This differs from many probabilistic accounts of generics which attempt to account for generics using real-world frequency (Cohen 1996; Kochari, R. v. Rooij, and Schulz 2020; R. v. Rooij and Schulz 2020). For example, generics seem to be licenced when a property is particularly common among one category relative to alternative categories. There does seem to be a connection between these probability measures and generics, but these connections are *externalistic*. These measures describe facts about the real world which lead concepts to have the probability distributions they do, rather than describing the calculations that people make when speaking.

And so, proximately, the reason (24) is false is that there simply aren't enough paperback exemplars sampled to make (24) true. Likewise, someone can endorse (21) yet not (22), simply because the points sampled are not sufficiently female yet are sufficiently egg-layers. Ultimately, however, the probability distribution is shaped this way to ensure that we remember the *important* parts of a concept.

Egg-laying non-females may seem contradictory, but this is not the case. While it is true that egg-laying non-females do not exist (as far as I know, I am not an ornithologist), there may still be non-female egg-laying points. The fact there are such points, does not mean that the object that the point describes exists or one believes it to exist. For example, since $\mathscr{E}$-space is continuous, there are uncountably infinite points for any concept. We simply do not know of any *individual concepts* of ducks which are not female and that lay eggs.

4.3 Quantification over individuals and quantification over points in a concept

Similarly, one can believe (25) and agree that (24) is false.

(25) Most books are paperbacks.

The reason for this is that quantifiers like "every" or "three" *require* individual concepts to quantify over. Indeed, for many quantifiers, we need to be able to quantify over specific individuals rather than just regions in space. For example, I can say (26a) which goes over individuals, whereas I certainly cannot say (26b) which goes over points.

(26) a. Everyone came to the party: Matthew, Mark, Luke *and* John!

b. * Everyone came to the party: $\langle 2.6, 13.7\rangle$, $\langle 4.232, -23.1\rangle$ *and* $\langle 5.2122, 0.3\rangle$!

So, the reason we can assert (25) and still not believe the corresponding generic, is because "most" quantifies over the individual-concepts that are books, rather than simply any point in the book-manifold.

Leslie, Khemlani, and Glucksberg 2011 found that people sometimes endorse false statements quantified by "all" if the corresponding generic is true. There is a sharp distinction (for me), between (27a) and (27b). While (27a) is false, it seems considerably better than (27b).

(27) a. All ducks lay eggs.

b. Every duck lays eggs.

c. All of the ducks lay eggs.

One thing to note is that (27b) and (27c) can refer to a concrete set of ducks (e.g. the four ducks at the park). (27a), on the other hand, can only refer to ducks in general, i.e. our entire duck-manifold. One intriguing hypothesis is that "all" without a partitive, may actually quantify over points rather than individual concepts. Indeed, some analyses have previously connected non-partitive "all" to bare plural generics (Matthewson 2001). Leslie, Khemlani, and Glucksberg 2011 did not check if people endorsed generics prefaced with *every*. I suspect this would have considerably less endorsement if "every" quantifies over individual concepts, rather than points.

Chapter 5

Adjectives

Unlike nouns which denote (generally speaking) objects, adjectives describe properties of objects (Kennedy 2015). They extract some salient abstract property of a concept whether its height, beauty or even its “fakeness”.

Under certain analyses, these properties exist along a scale (Morzycki 2015; Kennedy and McNally 2005). For example, “tall” is determined based on an object’s position along a scale of height. These kinds of adjectives are known as gradable adjectives since they can be graded and compared. Not all adjectives are gradable, some adjectives only ascribe a specific property which cannot be compared. For example, (28a) has a gradable adjective whereas (28b) has a non-gradable adjective.

(28) a. John is taller than Mary.

b. ? John is more single than Mary.

Note that non-gradable adjectives can be coerced into comparatives.

(29) Jean is more French than Marie.

(29) can describe a scenario where Jean is more stereotypically French than Marie. Note that while there are many different approaches to dealing with gradable adjective, this work will largely take after a degrees based approach such as Cresswell 1976 or Kennedy and McNally 2005 with some modifications.

Many of the features proposed by prototype theory approaches to concepts *are* adjectives. Indeed, its easy to see how the scales used for adjectives are connected to the dimensions in featural conceptual theories. For example, size or weight could be such dimensions.

However, the $\mathscr{E}$-space does not explicitly model features as dimensions and so we can not directly model scales in the $\mathscr{E}$-space. Instead, adjectives simply implement a new conceptual space which models that scale and there is a *mapping* from the $\mathscr{E}$-space to the adjective space. Before describing the internal structure, we can already show some of the expressive power that Chapter 3 grants us for adjectives.

5.1 Adjectives simplified

Concretely, we will treat adjectives as a function from points to truth values, that take a concept as a free variable: Let $\lambda.\overrightarrow{\mathtt{x}_{\mathscr{E}}} A(C)(\overrightarrow{\mathtt{x}_{\mathscr{E}}})$ be an adjective used in a positive sense, i.e. without a comparative clause like in (30).

(30) John is skilled.

Some adjectives' meaning changes depending on the nouns they modify. For example, a skilled baker who is also a guitarist is not necessarily a skilled guitarist. We generalise to the worst case and assume that A has an additional argument, C so that the measurement would change depending on the concept given. Whether or not all adjectives have this extra parameter is a matter of open debate (Morzycki 2015) and is out of the scope of this work. The internal structure of A will be discussed in Section 5.2.

So, (30) might look something like this:

$$[\![\text{John is skilled}]\!] = \left(\mu \overrightarrow{\mathtt{j}_{\mathscr{E}}} \in John\right) Skilled(C)(\overrightarrow{\mathtt{j}_{\mathscr{E}}}) \tag{5.1}$$

C would naturally be filled in with whatever John was skilled at.

Sometimes, the adjective and the noun modified might not have any intersection. For example, there are no lions which are made of plastic.

$$[\![\text{plastic lion}]\!] = Plastic(C)(\overrightarrow{\mathtt{x}_{\mathscr{E}}}) \wedge Lion(\overrightarrow{\mathtt{x}_{\mathscr{E}}}) \tag{5.2}$$

There is no point which satisfies 5.2 since no lion is plastic, thus, by simple pragmatic reasoning we are *forced* to use the metaphorical operator, if we assume our interlocutor does not refer to empty extensions (this assumption is simply the Non-Vacuity Principle proposed by (Kamp and Partee 1995)). So, instead we need to interpret "lion" with the metaphoric operator.

$$[\![\text{plastic lion}]\!] = \lambda \overrightarrow{\mathtt{x}_{\mathscr{E}}}.Plastic(C)(\overrightarrow{\mathtt{x}_{\mathscr{E}}}) \wedge (Lion \upharpoonright Appearance)(\overrightarrow{\mathtt{x}_{\mathscr{E}}}) \tag{5.3}$$

The most salient things that satisfy this interpretation are, of course, little plastic toys which resemble lions!

5.1.1 Privative adjectives

Our story for privative adjectives is similar. A privative adjective then, is any adjective where $A(C)(\overrightarrow{\mathtt{x}_{\mathscr{E}}})$ is 0 whenever $\overrightarrow{\mathtt{x}_{\mathscr{E}}} \in C$. For example, when $\llbracket \text{Fake} \rrbracket = \lambda \overrightarrow{\mathtt{x}_{\mathscr{E}}}.Fake(Gun)(\overrightarrow{\mathtt{x}_{\mathscr{E}}})$, every point on the gun-manifold has $Fake(Gun)(\overrightarrow{\mathtt{x}_{\mathscr{E}}}) = 0$.

So, any privative adjective, A, meets the following condition:

$$F(C)(\overrightarrow{\mathtt{x}_{\mathscr{E}}}) = \begin{cases} 0 & \text{if } \overrightarrow{\mathtt{x}_{\mathscr{E}}} \in M_C \\ F(C)(\overrightarrow{\mathtt{x}_{\mathscr{E}}}) & \text{otherwise} \end{cases} \tag{5.4}$$

Concretely, let's look at the case of "fake gun".

$$\llbracket \text{fake gun} \rrbracket = \lambda \overrightarrow{\mathtt{x}_{\mathscr{E}}}.Fake(Gun)(\overrightarrow{\mathtt{x}_{\mathscr{E}}}) \wedge (Gun \upharpoonright C)\,(\overrightarrow{\mathtt{x}_{\mathscr{E}}})$$

Just like "plastic lion", there is no point in the gun-manifold which satisfies $Fake(Gun)$, so we use the metaphorical operator. So, a fake gun is something that is sufficiently similar to a gun in some respect to be called a gun, while also being fake with respect to guns. While the specific facts that define fake for fake gun probably relate to the purpose and appearance of the object (Del Pinal 2015) or the intent to fool someone (Guerrini 2018), *compositionally*, we say that it is something which is close to a gun, and has the property of fakeness with regards to guns (which precludes the possibility of it actually being a gun). Unlike "plastic lion", this privativity is an intrinsic part of the word "fake", so we could *never* find a fake gun that is *literally* a gun.

5.1.2 Fake fake guns

We also need to be able to recursively apply "fake" or other privatives (Martin 2019; Guerrini 2021). A "fake fake gun" is something that resembles a fake gun, but is not, in fact, a fake gun. To do this, we simply use our manifold coercion operator, $\ominus$ to turn "fake gun" into a concept triple. This now gives us a new concept we can treat like any other, consisting of fake guns. Our new conceptual-triple (which we'll call Fg for brevity) is $Fg = \ominus(C, Fake(Gun)(\overrightarrow{\mathtt{x}_{\mathscr{E}}}) \wedge (Gun \upharpoonright C)\,(\overrightarrow{\mathtt{x}_{\mathscr{E}}}))$ which will select anything that met our definition for fake gun.[1] After that, we can just use fake like we did before.

$$\llbracket \text{fake fake gun} \rrbracket = \lambda \overrightarrow{\mathtt{x}_{\mathscr{E}}}.Fake(Fg)(\overrightarrow{\mathtt{x}_{\mathscr{E}}}) \wedge (Fg \upharpoonright C)\,(\overrightarrow{\mathtt{x}_{\mathscr{E}}})$$

Crucially, *real* guns which resemble fake guns can meet this specification! This is because fake guns are similar to real guns, therefore things which are

[1] Since everything that satisfies $(Gun \upharpoonright C)$ must also be on the manifold of C. Otherwise, we couldn't measure how similar it was with respect to C.

similar to fake guns will include real guns. This could, in principle, be continued recursively indefinitely, although on a performance level, even "fake fake fake gun" is very hard to understand.

5.1.3 Fake metaphors

Guerrini and Mascarenhas 2019 note that privatives can *also* apply to metaphorical uses directly. For example, a professional wrestler could be considered a lion in terms of his prowess as a fighter (e.g. $(Lion \upharpoonright Fighter)$). Now, let's say we discovered this man had merely been pretending to fight, and in fact, was a pretty bad fighter. We could then say he was a "fake lion"!

We can treat this metaphorical usage the exact same way we treated "Fake fake", by applying $\ominus$ to our metaphorical use.

$$\begin{aligned}[\![\text{fake lion}]\!] = \lambda \overrightarrow{\mathrm{x}_{\mathscr{E}}}.Fake(\ominus(Fighter, (Lion \upharpoonright Fighter)))(\overrightarrow{\mathrm{x}_{\mathscr{E}}})\\ \wedge (\ominus(Fighter, (Lion \upharpoonright Fighter)) \upharpoonright C)\,(\overrightarrow{\mathrm{x}_{\mathscr{E}}})\end{aligned}$$

The wrestler *resembles* a lion-in-terms-of-fighting, but is not, in fact, a lion-in-terms-of-fighting. The C we compare along could be $Appearance$, or perhaps in this scenario, $Performer$. This is because the wrestler *performs* like someone who is a lion in terms of fighting.

5.1.4 Fake individuals

Privative adjectives don't just apply to categories, but individuals too! For example, a deep-fake video featured a "fake Obama" and a video game featured a type of enemy called "fake Hitler". We can even have contingent privative modification of individuals: "[a]n image of a White Barack Obama".

Handling these cases is remarkably easy, we simply coerce our individual concepts to be category concepts and proceed as normal.

$$[\![\text{fake Obama}]\!] = Fake(Obama)(\overrightarrow{\mathrm{x}_{\mathscr{E}}}) \wedge (Obama \upharpoonright Appearance)\,(\overrightarrow{\mathrm{x}_{\mathscr{E}}}) \qquad (5.5)$$

Fake Obama is fake with regard to his Obama-ness yet is sufficiently Obama-like to be referred to as Obama.

5.2 Adjective spaces

Of course, the *internal* structure of adjectives is important. They are not simple category manifolds since some are gradable and have a comparative form. This section defines a degree-based structure for adjectives. Furthermore, it can be easily added to the computational model described in Section 3.6.

For any adjective, A, we define a map $f_{\mathsf{A}}(C) : M_A \rightarrow \mathsf{A}$. $f_{\mathsf{A}}(C)$ is a function which maps points from a subset of $M_A \subseteq \mathscr{E}$ to points in A, a conceptual-space defined for the adjective, A. M_A consists of the points in $\mathscr{E}$ for which the adjective is defined. C is the concept used to model subsective adjectives.

A is a new conceptual space which differs from $\mathscr{E}$ as it is used to define a *partial order*. Specifically, we define the ordering, $\underset{\mathsf{A}}{\preceq}$ over all the individuals of A. Contrary to many degree based accounts, we do not assume the underlying scales are a total order (Kennedy 2007). Assuming a total order for all adjectives forces us to draw certain entailments that are not necessarily valid. For instance, it is unclear that "beauty" can be arranged in a single linear order, even when restricting beauty to a single category.

A simple model that may provide a decent first approximation is Order Embeddings (Vendrov et al. 2016). Concretely, Order Embeddings define a partial order embedding in $\mathbb{R}^n$. Given two vectors, $\vec{x}$ and $\vec{y}$ in $\mathbb{R}^n$, $\vec{x} \preceq \vec{y}$ if and only if $\vec{x}_i \leq \vec{y}_i$ for $1 \leq i \leq n$. It is easy to see that this reduces to a total order if $n = 1$. Adjectives might share the same embeddings, or map embeddings to similar but modified embeddings. For instance, tall and short clearly have a related conceptual space, but simply inversed.

5.2.1 Comparatives

As with degree-based approaches, the comparative has a relatively straightforward semantics (McNally 2016).

$$\lambda \overrightarrow{\mathrm{x}_{\mathscr{E}}}.\lambda \overrightarrow{\mathrm{y}_{\mathscr{E}}}.\, [\![\overrightarrow{\mathrm{x}_{\mathscr{E}}} \text{ is taller than } \overrightarrow{\mathrm{y}_{\mathscr{E}}}]\!] = f_{\mathsf{Tall}}(\overrightarrow{\mathrm{y}_{\mathscr{E}}}) \underset{\mathsf{Tall}}{\preceq} f_{\mathsf{Tall}}(\overrightarrow{\mathrm{x}_{\mathscr{E}}})$$

$\overrightarrow{\mathrm{x}_{\mathscr{E}}}$ is taller than $\overrightarrow{\mathrm{y}_{\mathscr{E}}}$ if each dimension of $f_{\mathsf{Tall}}(\overrightarrow{\mathrm{x}_{\mathscr{E}}})$ is greater than $f_{\mathsf{Tall}}(\overrightarrow{\mathrm{y}_{\mathscr{E}}})$, in the new conceptual space, Tall.

To compare actual individuals, the scenario is a bit different.

$$[\![\text{taller than Bill}]\!] = \lambda \overrightarrow{\mathrm{x}_{\mathscr{E}}}.S\left(\left(\mu \overrightarrow{\mathrm{b}_{\mathscr{E}}} \in Bill\right) f_{\mathsf{Tall}}(\overrightarrow{\mathrm{b}_{\mathscr{E}}}) \underset{\mathsf{Tall}}{\preceq} f_{\mathsf{Tall}}(\overrightarrow{\mathrm{x}_{\mathscr{E}}}), \alpha\right)$$

Concretely, we iterate over the exemplars of Bill with μ and see whether $\overrightarrow{\mathrm{x}_{\mathscr{E}}}$ is greater than that exemplar. This gives us the *percentage* of sampled Bills that $\overrightarrow{\mathrm{x}_{\mathscr{E}}}$ is taller than. So, we need to wrap the whole thing in a step function in order to make the truth value binary. If we set $\alpha = 0.95$, we could ensure that something is taller than Bill when it is taller than roughly 95% of its exemplars.

5.2.2 Positive use

To use a gradable adjective without a comparative (positive form), we use a very similar structure to comparatives. We cannot compare to a single point like commonly done in degree semantics because we use a partial order. Instead, we compare $\overrightarrow{\mathrm{x}_{\mathscr{E}}}$ to all sampled exemplars. Let *Comp* be a contextually chosen-manifold in $\mathscr{E}$ which serves as a standard of comparison for our adjective, *A*. For example, *Comp* would typically be *People* when we describe a person as tall.

$$[\![\text{is A}]\!] = \lambda A.\lambda \overrightarrow{\mathrm{x}_{\mathscr{E}}}.S\left((\mu \overrightarrow{\mathrm{y}_{\mathscr{E}}} \in Comp)\, f_{\mathsf{A}}(\overrightarrow{\mathrm{y}_{\mathscr{E}}}) \underset{\mathsf{A}}{\prec} f_{\mathsf{A}}(\overrightarrow{\mathrm{x}_{\mathscr{E}}}), \alpha\right)$$

We then calculate the percentage of samples from *Comp* that $\overrightarrow{\mathrm{x}_{\mathscr{E}}}$ is greater than. We can then determine the standard at which $\overrightarrow{\mathrm{x}_{\mathscr{E}}}$ is compared to *Comp* by setting This could very lexically like with Kennedy and McNally 2005 the appropriate adjective. Concretely, when the standard is max for Kennedy, we use $\alpha = 1$.

5.2.3 Absolute gradable adjectives and measure individuals

This works for *relative* adjectives where the standard is defined compared to other exemplars. Other gradable adjectives are *absolute*, the standard is independent of its comparisons; a full glass is not full because it is more full than some alternatives. For these, we can simply compare directly to the maximum of the scale.

$$[\![\text{is A}_{Abs}]\!] = \lambda A.\lambda \overrightarrow{\mathrm{x}_{\mathscr{E}}}.\max(\mathsf{A}) \underset{\mathsf{A}}{\preceq} f_{\mathsf{A}}(\overrightarrow{\mathrm{x}_{\mathscr{E}}}) \tag{5.6}$$

So, a full glass is full if and only if it is as full or fuller than the maximally full thing.

Rather than having a different interpretation for absolute adjectives, we define individuals in $\mathscr{E}$ which map to the maximum of A and then use the same interpretation for both. For example, the individual concept for a metre, would map to a position in both the height and length spaces which is equivalent to one metre. Likewise, we do this with concepts such as half-full, full, or empty. Concretely, *Comp* for an absolute adjective is an individual concept, rather than a category concept. For example, for *Full*, *Comp* would be set to some individual *F* such that $f_{\mathsf{Full}}(F) = \max(\mathsf{Full}) - \varepsilon$. Note that we have to subtract some infinitesimally small quantity, ε from the maximum because we use $\underset{\mathsf{Full}}{\prec}$ rather than $\underset{\mathsf{Full}}{\preceq}$ in the normal interpretation.

$$[\![\text{is full}]\!] = \lambda \overrightarrow{\mathrm{x}_{\mathscr{E}}}.S\left((\mu \overrightarrow{\mathrm{y}_{\mathscr{E}}} \in F)\, f_{\mathsf{Full}}(\overrightarrow{\mathrm{y}_{\mathscr{E}}}) \underset{\mathsf{Full}}{\prec} f_{\mathsf{Full}}(\overrightarrow{\mathrm{x}_{\mathscr{E}}}), \alpha\right) \tag{5.7}$$

This equation reduces to equation 5.6, since $f_{\mathsf{Full}}(\overrightarrow{\mathrm{y}_{\mathscr{E}}}) = \max(\mathsf{Full})$ for $\overrightarrow{\mathrm{y}_{\mathscr{E}}} \in F$.

$$\begin{aligned}
[\![\text{is full}]\!] &= \lambda \overrightarrow{\mathrm{x}_{\mathscr{E}}}.S\left((\mu \overrightarrow{\mathrm{y}_{\mathscr{E}}} \in F)\, f_{\mathsf{Full}}(\overrightarrow{\mathrm{y}_{\mathscr{E}}}) \underset{\mathsf{Full}}{\prec} f_{\mathsf{Full}}(\overrightarrow{\mathrm{x}_{\mathscr{E}}}), \alpha \right) \\
&= \lambda \overrightarrow{\mathrm{x}_{\mathscr{E}}}.S\left((\mu \overrightarrow{\mathrm{y}_{\mathscr{E}}} \in F) \max(\mathsf{A}) - \varepsilon \underset{\mathsf{Full}}{\prec} f_{\mathsf{Full}}(\overrightarrow{\mathrm{x}_{\mathscr{E}}}), \alpha \right) \\
&= \lambda \overrightarrow{\mathrm{x}_{\mathscr{E}}}.S\left((\mu \overrightarrow{\mathrm{y}_{\mathscr{E}}} \in F) \max(\mathsf{A}) \underset{\mathsf{Full}}{\preceq} f_{\mathsf{Full}}(\overrightarrow{\mathrm{x}_{\mathscr{E}}}), \alpha \right) \\
&= \lambda \overrightarrow{\mathrm{x}_{\mathscr{E}}}.S\left(\max(\mathsf{A}) \underset{\mathsf{Full}}{\preceq} f_{\mathsf{Full}}(\overrightarrow{\mathrm{x}_{\mathscr{E}}}), \alpha \right) \\
&= \lambda \overrightarrow{\mathrm{x}_{\mathscr{E}}}. \max(\mathsf{A}) \underset{\mathsf{Full}}{\preceq} f_{\mathsf{Full}}(\overrightarrow{\mathrm{x}_{\mathscr{E}}})
\end{aligned}$$

Furthermore, since these measures exist in $\mathscr{E}$, we can compare measure items in terms of similarity, typicality or any other thing we might do in $\mathscr{E}$. For example, all measure individuals corresponding to men's heights will lie on the men's height manifold, and we can thus determine a typical height for a man using the same tools we had before.

5.2.4 Explicitly privative adjectives

To revisit our definition of privative adjectives from Section 5.1.1, recall that a privative adjective, $PrivAdj$, is any adjective which satisfies the following condition:

$$PrivAdj(C)(\overrightarrow{\mathrm{x}_{\mathscr{E}}}) = \begin{cases} 0 & \text{if } \overrightarrow{\mathrm{x}_{\mathscr{E}}} \in C \\ PrivAdj(C)(\overrightarrow{\mathrm{x}_{\mathscr{E}}}) & \text{otherwise} \end{cases} \tag{5.8}$$

With our new representation of adjectives, we can describe this a bit more explicitly:

$$f_{\mathsf{PrivAdj(C)}}(\overrightarrow{\mathrm{x}_{\mathscr{E}}}) = \begin{cases} f_{\mathsf{PrivAdj(C)}} & \text{if } \overrightarrow{\mathrm{x}_{\mathscr{E}}} \notin C \\ \min(\mathsf{PrivAdj}) & \text{if } \overrightarrow{\mathrm{x}_{\mathscr{E}}} \in C \end{cases} \tag{5.9}$$

Since the positive use of an adjective uses $\underset{\mathsf{F}}{\prec}$ to compare the samples, any member of C will never have a truth value about 0, since each one is at the minimum.

Chapter 6

Reasoning

This work has developed a way where we can check the truth-value of a proposition in a particular $\mathscr{E}$-space. While this is useful for analyses of privatives or generics, it does not allow us to reason about propositions. To reason, we need to *change* the $\mathscr{E}$-space to accommodate this new information.

People do not live in a static world of constant unchanging knowledge. Rather, from novel information, we change our beliefs about the world. From this process of change, our conceptual space will not be fixed, wooden shapes, but rather a moving, living thing which will change over time.

This chapter puts forward a tentative system which can account for certain reasoning puzzles. Further work will be needed to elaborate and precise this account to make a fully dynamic system which can handle the kinds of inferences people make on a day to day basis. For one, there is no way to represent events in our system as it stands. Indeed, without a proper understanding of events and causality between events, a full theory of reasoning will be completely impossible.

The system is best suited for a dynamic approach where the model of our $\mathscr{E}$-space changes as a we receive novel information and consider alternatives. Since propositions are represented as the sets of all possible $\mathscr{E}$-spaces where they are true (c.f. Section 3.2.1), reasoning over these propositions directly would be completely computationally infeasible. Rather, we reason by considering alternative $\mathscr{E}$-spaces, updating them as we get new information, and then ultimately deciding between them. More specifically, I assume a version of the Erotetic Theory of Reasoning (Koralus and Mascarenhas 2013). $\mathscr{E}$-space allows us to address notions of typicality or similarity which are not possible in the original formalisation of ETR.

6.1 The Erotetic Theory of Reasoning and Alternatives

The Erotetic Theory of Reasoning can be broadly considered as a theory of reasoning where *questions* are central. Concretely, human reasoning works primarily around organising different alternatives of the ways the world could be (i.e. disjunctions of various propositions) and then trying to reduce the set of these different alternatives. The system is fully formalised and can be applied to explain various fallacious patterns in human reasoning, but also the correct inferences that people make robustly. The formalised system, however, does make certain assumptions that make notions of typicality or similarity difficult to handle.

Often, we learn something that strongly implicates something, without guaranteeing it. Imagine I tell you that Jim is either a lion or a zebra. I then tell you he has incredibly sharp teeth. It is of course, possible for zebra to have sharp teeth, but it is certainly more associated with lions than with zebras. While it does not logically follow in either classical logic or the first formalisation of ETR that Jim is a lion, I am rather certain that people would make this conclusion, precisely because lions typically eat gazelles whereas zebras do not.

6.2 Representing alternatives

We need a way to represent alternatives if we are to reason about them. The solution is to give each alternative a full model of the world with $\mathscr{E}$ and any other conceptual spaces. The model has the same signature, since we assume the same concepts exist in both.

(31) John is French or Mary is English.

For example, given (31), we have two alternatives, η_1 and η_2. η_1 corresponds to the scenario where John is French, and η_2 the scenario where Mary is English. Concretely, this means that the individual concept for John, J is entirely on the French manifold in η_1 where as J is not necessarily French in η_2. [1]

Until now, we have been assuming there is only one model for our concepts. To keep track of which alternative η that a concept C is in, we write C^η to indicate that this is C in η. Likewise, φ^η would give us the truth-value of φ in η.

[1] There is a very interesting question of how individual concepts are instantiated for novel individuals, but I do not address it in this work. For example, it's likely that people would assume J lies on the man-manifold in both alternatives.

6.3 Monotonic update

To learn a new piece of information, we need to update our alternatives. Our update is *monotonic*; it does not account for the complicated matter of revising beliefs. Like a function which is only increasing (and thus monotonic), our information only increases with new information; we never walk back on our own beliefs.

Our update function takes a concept and a new property about that concept and returns an updated version of that concept, in a given model. Specifically, it restricts a concept to a subset of that concept which has the property in question.

Let C be the concept to be updated. Let $\varphi(\overrightarrow{\mathrm{x}_{\mathscr{E}}})$ be an open proposition which is the new property about C. For example, when our new premise is $(\mu\overrightarrow{\mathrm{x}_{\mathscr{E}}} \in C)\, \varphi(\overrightarrow{\mathrm{x}_{\mathscr{E}}})$, we apply φ to the C-individual. We use square bracket notation to indicate our update function; $\eta\,[C,\varphi]$ is an alternative, η where we have updated C^{η} with a property, φ. We can apply this continuously, for example, $\eta\,[C,\varphi]\,[D,\psi]$ would then update that updated model.

$$\eta\,[C,\varphi] = \eta \text{ except that } C^{\eta[C,\varphi]} = \ominus(C^{\eta},\varphi^{\eta}) \tag{6.1}$$

The concept coercion operator, $\ominus$ (c.f. Section 3.2.7) takes C and returns the submanifold of C where $\varphi(\overrightarrow{\mathrm{x}_{\mathscr{E}}})$ is true. The new alternative, $\eta\,[C,\varphi]$ is thus η where we have restricted C to the parts where $\varphi(\overrightarrow{\mathrm{x}_{\mathscr{E}}})$ is true.

While this update does not contradict our previous model (the manifold can only shrink), the propositions we assent to may change. For example, if the probability mass of a concept J^{η} was *mostly* on the manifold for a predicate P^{η}, then we would assent to $(\mu\overrightarrow{\mathrm{x}_{\mathscr{E}}} \in J)\, P(\overrightarrow{\mathrm{x}_{\mathscr{E}}})$, even if M_J^{η} was not a subset of M_P^{η}. This is because most of the sampled exemplars would lie on P^{η}, even if the entire manifold did not. In the alternative where we update with $\neg P(J)$, the individual, $J^{\eta[J,\neg P(\overrightarrow{\mathrm{x}_{\mathscr{E}}})]}$ would be the part of J^{η} that was *not* on P^{η}. The manifold shrank monotonically, but the propositions we assent to did change, since now no points of $J^{\eta[J,\neg P(\overrightarrow{\mathrm{x}_{\mathscr{E}}})]}$ lay on $P^{\eta[J,\neg P(\overrightarrow{\mathrm{x}_{\mathscr{E}}})]}$.

If there are more concepts involved, things get trickier. For example, say we learn that (32) and we wish to update our beliefs.

(32) John is taller than Mary.

We need to decide whether to update John or Mary (or both). One intriguing hypothesis is that we *center* our propositions around some topic (Bittner 2007). In other words, we choose whether to update η to $\eta[J, TallerThanMary(\overrightarrow{\mathrm{x}_{\mathscr{E}}})]$ or $\eta[M, JohnIsTallerThan(\overrightarrow{\mathrm{x}_{\mathscr{E}}})]$ based on which individual is more important contextually.

We can update John, but then we still stochastically compare to Mary. If some candidate Mary is very tall, yet has very low probability of being sampled, we may not ensure that each John is taller than this Mary. However, we still would assent to (32) after the update since those tall Mary-exemplars are so infrequently sampled.

6.4 Deciding between alternatives

Ultimately, what we really want is a way to reach a conclusion.

One simple situation where we want to eliminate alternatives is if there is a contradiction. For example, given the same scenario in (31), we have two alternatives η_1, η_2 where $French(John)$ in η_1 and $English(Mary)$ in η_2.

If we then learn that Mary is *not* English, and apply the monotonic update to each alternative, Mary in η_2 will shrink to nothing. This is because there will be no part of her manifold which is not on the English manifold, so we are forced to reduce her to an empty manifold. We cannot sample points from an empty manifold, so this describes a contradictory concept! As a result, we eliminate η_2 from our set of alternatives. With only η_1 remaining, we conclude that John is French since he is French in η_1, giving us the disjunctive syllogism.

Unfortunately, there are many scenarios where we *do not* have only a single alternative left, yet we want to draw a conclusion. In these cases, we introduce a parameter γ which is a threshold of truth for $\varphi^{\eta}(C^{\eta})$. Recall $\varphi(C)$ is short form for $(\mu \overrightarrow{\mathrm{x}_{\mathscr{E}}} \in C)\, \varphi(\overrightarrow{\mathrm{x}_{\mathscr{E}}})$. If an alternatives assigns a truth-value to our novel information, which is lower than γ, then we strike that η from our alternatives. This is like the Q-Update in the original formalisation of ETR (Koralus and Mascarenhas 2013).

Let $H = \{\eta_1, \eta_2, \ldots\}$ be the set of our alternative models.

$$H[C, \varphi, \gamma] = \left\{ \eta[C, \varphi] \mid \forall \eta \in H \text{ such that } M_{C^{\eta[C,\varphi]}} \neq \varnothing \text{ and } \varphi^{\eta}(C^{\eta}) \geq \gamma \right\} \quad (6.2)$$

$H[C, \varphi, \gamma]$ is the set of alternatives where each alternative has been updated with C and φ. It also excludes alternatives where C^{η} does not get updated with an empty manifold, and where $\varphi(C^{\eta})^{\eta}$ at least as true as γ.

If we are forced to draw a conclusion, we would want the $\eta \in H$ that has the maximum likelihood before the update, just like in the original ETR. To do this, we could do $H[C, \varphi, \max_{\eta \in H}(\varphi^{\eta}(C^{\eta}))]$ to force ourselves to choose the most likely alternative by setting γ to whatever is the highest truth-value for $\varphi(C)$ across all alternatives.

6.5 Illusory Inferences from Disjunction

Illusory Inferences from Disjunction (IIFDs) are a class of fallacies which people have been robustly shown to make which can be accounted for by the Erotetic Theory of Reasoning (Walsh and P. N. Johnson-Laird 2004; Khemlani and P. N. Johnson-Laird 2009; Koralus and Mascarenhas 2013). While it may seem to have little to do with reasoning by representativeness, its structural form is similar to the conjunction fallacy (Sablé-Meyer and Mascarenhas 2021) which is the most classical case of reasoning by representativeness. The fallacy is shown in Table 6.1.

John is English or else Mary is French and James is German.
James is German.

$\therefore$ Mary is French

Table 6.1: An example of the illusory inference from disjunction.

We can account for it in our modified version of the Erotetic Theory of Reasoning as well. Concretely, we generate two sets of alternatives: $\eta = \{\eta_1, \eta_2\}$ where $English(John)^{\eta_1}$, $French(Mary)^{\eta_2}$, and $German(James)^{\eta_2}$.

We perform the monotonic update with our second premise; "James is German". $James^{\eta_2}$ will not change at all since he was already German. $James^{\eta_1}$ however, will change (and now be equivalent to $James^{\eta_2}$). Thus $\llbracket \text{James is German} \rrbracket^{\eta_2} = 1$ whereas $\llbracket \text{James is German} \rrbracket^{\eta_1} = \varepsilon$ where ε is the default truth-value that an individual is German. ε is almost certainly a small number, and therefore, as long as $\gamma \geq \varepsilon$, we will eliminate the first set of alternatives via the update rule in 6.2. Furthermore, we likely choose $\gamma = \max_{\eta \in H}(\llbracket \text{James is German} \rrbracket^{\eta})$ since we are forced to make a conclusion.

Indirect IIFDs are a more general case of IIFD noted by Sablé-Meyer and Mascarenhas 2021 where the second premise does not match any part of first premise, but is *associated* with part of the first premise, as shown in Table 6.2.

John is English or else Mary is French and James is German.
James loves sauerkraut.

$\therefore$ Mary is French

Table 6.2: An example of the indirect illusory inference from disjunction.

Unlike traditional IIFDs, this variant cannot be accounted by the original formalisation of the erotetic theory of reasoning nor simple premise-matching accounts. "James loves sauerkraut" is associated with "James is German" (since Germans are more typically assumed to love sauerkraut more than non-Germans).

We initiate the same alternatives $\eta = \{\eta_1, \eta_2\}$ where $English(John)^{\eta_1}$, $French(Mary)^{\eta_2}$, and $German(James)^{\eta_2}$. When we perform the monotonic update, we do the exact same procedure. The only difference is that the additional premise "James loves sauerkraut" has a much lower truth-value in η_2 than "James is German" did. Nevertheless, it's our only premise and we are forced to make a conclusion. We conclude Mary is French because James loving sauerkraut has a higher truth-value if we know he is German than if we do not.

$$\varphi(\overrightarrow{\mathrm{x}_{\mathscr{E}}}) = \lambda \overrightarrow{\mathrm{x}_{\mathscr{E}}}. [\![\text{loves sauerkraut}]\!] (\overrightarrow{\mathrm{x}_{\mathscr{E}}})$$

$$H = \{\eta_1, \eta_2\}$$

$$H[James, \varphi, \max_{\eta \in H}([\![\text{James is German}]\!]^{\eta})] = \{\eta_2[James, \varphi]\}$$

6.6 The Conjunction Fallacy

The conjunction fallacy is a classic example of fallacious reasoning where people wrongly assert that $P(a) < P(a \cap b)$, despite that violating the basic axioms of probability (Tversky and Kahneman 1974). The classic example is the following question:

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.
>
> Which is more probable?
>
> 1) Linda is a bank teller.
> 2) Linda is a bank teller and is active in the feminist movement.

People robustly prefer 2) to 1) despite this violating basic probability. Furthermore, Tversky and Kahneman have shown that this is not simply the result of pragmatically strengthening 1 to mean "Linda is a bank teller and not active in the feminist movement". Sablé-Meyer and Mascarenhas 2021 argue that the conjunction fallacy should be viewed as a special case of an indirect IIFD.

Concretely, we view the scenario as deciding between two different alternatives, η_1 and η_2 where $BankTeller(L)^{\eta_1}$ and $(BankTeller(L) \wedge Feminist(l))^{\eta_2}$. Our φ in this case is the little preamble about Linda, which will naturally lead to $[\![\text{Linda's description}]\!]^{\eta_2} > [\![\text{Linda's description}]\!]^{\eta_1}$ since the preamble is much more indicative of a feminist than of someone who may or may not be a feminist. Since we are forced by the question to draw a conclusion, we set γ to be the higher of the two and so we conclude η_2 is the right alternative model.

Bibliography

[1] Nicholas Asher and Michael Morreau. "What Some Generic Sentences Mean". In: *The Generic Book*. Ed. by Greg N. Carlson and Francis Jeffry Pelletier. University of Chicago Press, 1995, pp. 300–339.

[2] Marco Baroni. "Grounding Distributional Semantics in the Visual World". In: *Language and Linguistics Compass* 10.1 (2016), pp. 3–13. ISSN: 1749-818X. DOI: https://doi.org/10.1111/lnc3.12170. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/lnc3.12170 (visited on 05/06/2021).

[3] Marco Baroni, Raffaella Bernardi, and Roberto Zamparelli. "Frege in Space: A Program for Compositional Distributional Semantics". In: *Linguistic Issues in Language Technology* 9 (Jan. 1, 2014). ISSN: 1945-3604. DOI: 10.33011/lilt.v9i.1321. URL: https://journals.colorado.edu/index.php/lilt/article/view/1321 (visited on 11/30/2021).

[4] Lawrence Barsalou. "Perceptual symbol systems". In: *The Behavioral and Brain Sciences* 22.4 (Aug. 1999), 577–609, discussion 610–660. ISSN: 0140-525X. DOI: 10.1017/s0140525x99002149.

[5] Jens Behrmann et al. *Invertible Residual Networks*. arXiv:1811.00995. arXiv, May 18, 2019. DOI: 10.48550/arXiv.1811.00995. arXiv: 1811.00995[cs,stat]. URL: http://arxiv.org/abs/1811.00995 (visited on 05/25/2022).

[6] Sudeep Bhatia. "Associative judgment and vector space semantics." In: *Psychological Review* 124.1 (Jan. 2017), pp. 1–20. ISSN: 1939-1471, 0033-295X. DOI: 10.1037/rev0000047. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/rev0000047 (visited on 06/04/2022).

[7] Sudeep Bhatia and Ada Aka. "Cognitive Modeling With Representations From Large-Scale Digital Data". In: *Current Directions in Psychological Science* (Apr. 6, 2022), p. 096372142110681. ISSN: 0963-7214, 1467-8721. DOI: 10.1177/09637214211068113. URL: http://journals.sagepub.com/doi/10.1177/09637214211068113 (visited on 06/04/2022).

[8] Maria Bittner. "Online Update: Temporal, Modal, and de Se Anaphora in Polysynthetic Discourse". In: *Direct Compositionality*. Ed. by Chris Barker and Pauline Jacobson. Oxford University Press, 2007, pp. 11–363. URL: https://philarchive.org/rec/BITOUT (visited on 06/03/2022).

[9] Gemma Boleda. "Distributional Semantics and Linguistic Theory". In: *Annual Review of Linguistics* 6.1 (2020), pp. 213–234. DOI: 10.1146/annurev-linguistics-011619-030303. URL: https://doi.org/10.1146/annurev-linguistics-011619-030303.

[10] Susan Carey. *The Origin of Concepts*. New York: Oxford University Press, 2009. 608 pp. ISBN: 978-0-19-536763-8. DOI: 10.1093/acprof:oso/9780195367638.001.0001. URL: https://oxford.universitypressscholarship.com/10.1093/acprof:oso/9780195367638.001.0001/acprof-9780195367638 (visited on 11/23/2021).

[11] Gregory Norman Carlson. "Reference to Kinds in English". In: *Doctoral Dissertations Available from Proquest* (Jan. 1, 1977), pp. 1–506. URL: https://scholarworks.umass.edu/dissertations/AAI7726414.

[12] Emmanuel Chemla, Brian Buccola, and Isabelle Dautriche. "Connecting Content and Logical Words". In: *Journal of Semantics* 36.3 (Aug. 2019), pp. 531–547. DOI: 10.1093/jos/ffz001. URL: https://hal.archives-ouvertes.fr/hal-02474320 (visited on 12/07/2021).

[13] Emmanuel Chemla, Isabelle Dautriche, et al. "Constraints on the lexicons of human languages have cognitive roots present in baboons (Papio papio)". In: *Proceedings of the National Academy of Sciences* 116.30 (July 23, 2019). Publisher: Proceedings of the National Academy of Sciences, pp. 14926–14930. DOI: 10.1073/pnas.1907023116. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1907023116 (visited on 06/04/2022).

[14] Ariel Cohen. "On the Generic Use of Indefinite Singulars". In: *Journal of Semantics* 18.3 (2001). Publisher: Oxford University Press, pp. 183–209. DOI: 10.1093/jos/18.3.183.

[15] Ariel Cohen. "Think generic! the meaning and use of generic sentences". PhD thesis. USA: Carnegie Mellon University, 1996. 225 pp.

[16] Ariel Cohen. *Think generic! the meaning and use of generic sentences*. Dissertations in linguistics. Stanford, Calif: CSLI, 1999. 208 pp. ISBN: 978-1-57586-207-1 978-1-57586-208-8.

[17] Alexis Conneau et al. "Word Translation Without Parallel Data". In: *arXiv:1710.04087 [cs]* (Jan. 30, 2018). arXiv: 1710.04087. URL: http://arxiv.org/abs/1710.04087 (visited on 05/05/2022).

[18] Keenan Crane et al. "A Survey of Algorithms for Geodesic Paths and Distances". In: *arXiv:2007.10430 [cs]* (July 20, 2020). arXiv: 2007.10430. URL: http://arxiv.org/abs/2007.10430 (visited on 04/27/2022).

[19] M. J. Cresswell. "The Semantics of Degree". In: *Montague Grammar*. Ed. by Barbara Partee. Academic Press, Jan. 1, 1976, pp. 261–292. ISBN: 978-0-12-545850-4. DOI: 10.1016/B978-0-12-545850-4.50015-7. URL: https://www.sciencedirect.com/science/article/pii/B9780125458504500157 (visited on 05/17/2022).

[20] Simon De Deyne, Amy Perfors, and Daniel J Navarro. "Predicting human similarity judgments with distributional models: The value of word associations." In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. COLING 2016. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 1861–1870. URL: https://aclanthology.org/C16-1175 (visited on 06/04/2022).

[21] Guillermo Del Pinal. "Dual Content Semantics, privative adjectives, and dynamic compositionality". In: *Semantics and Pragmatics* 8.0 (Mar. 27, 2015), pp. 7–53. ISSN: 1937-8912. DOI: 10.3765/sp.8.7. URL: https://semprag.org/index.php/sp/article/view/sp.8.7 (visited on 11/23/2021).

[22] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv:1810.04805 [cs]* (May 24, 2019). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805 (visited on 05/08/2021).

[23] James J. DiCarlo, Davide Zoccolan, and Nicole C. Rust. "How Does the Brain Solve Visual Object Recognition?" In: *Neuron* 73.3 (Feb. 9, 2012), pp. 415–434. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2012.01.010. URL: https://www.sciencedirect.com/science/article/pii/S089662731200092X (visited on 02/23/2022).

[24] Laurent Dinh, David Krueger, and Yoshua Bengio. *NICE: Non-linear Independent Components Estimation*. Apr. 10, 2015. arXiv: 1410.8516[cs]. URL: http://arxiv.org/abs/1410.8516 (visited on 05/25/2022).

[25] Allyson Ettinger. "What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models". In: *Transactions of the Association for Computational Linguistics* 8 (Dec. 2020), pp. 34–48. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00298. URL: https://direct.mit.edu/tacl/article/43535 (visited on 11/27/2021).

[26] Kit Fine. “Vagueness, Truth and Logic”. In: *Synthese* 30.3 (1975), pp. 265–300. ISSN: 0039-7857. URL: https://www.jstor.org/stable/20115033 (visited on 05/16/2022).

[27] Jerry Fodor. “The present status of the innateness controversy”. In: *RePresentations: Philosophical Essays on the Foundations of Cognitive Science*. Ed. by Jerry Fodor. MIT Press, 1981, pp. 257–316.

[28] Jerry Fodor and Ernest Lepore. *The compositionality papers*. Oxford University Press, 2002. ISBN: 0-19-925216-5.

[29] Jerry Fodor and Ernest Lepore. “The red herring and the pet fish: why concepts still can’t be prototypes”. In: *Cognition* 58.2 (Feb. 1, 1996), pp. 253–270. ISSN: 0010-0277. DOI: 10.1016/0010-0277(95)00694-X. URL: https://www.sciencedirect.com/science/article/pii/001002779500694X (visited on 11/23/2021).

[30] Bradley Franks. “Sense generation: A “quasi-classical” approach to concepts and concept combination”. In: *Cognitive Science* 19.4 (Oct. 1, 1995), pp. 441–505. ISSN: 0364-0213. DOI: 10.1016/0364-0213(95)90008-X. URL: https://www.sciencedirect.com/science/article/pii/036402139590008X (visited on 11/28/2021).

[31] Gottlob Frege. “Über Sinn Und Bedeutung”. In: *Zeitschrift für Philosophie Und Philosophische Kritik* 100.1 (1892), pp. 25–50.

[32] Juan A. Gallego et al. “Neural Manifolds for the Control of Movement”. In: *Neuron* 94.5 (June 7, 2017), pp. 978–984. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2017.05.025. URL: https://www.sciencedirect.com/science/article/pii/S0896627317304634 (visited on 05/23/2022).

[33] Octavian-Eugen Ganea, Gary Becigneul, and Thomas Hofmann. “Hyperbolic Neural Networks”. In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc., 2018. URL: https://papers.nips.cc/paper/2018/hash/dbab2adc8f9d078009ee3fa810bea142-Abstract.html (visited on 11/30/2021).

[34] Artur d’Avila Garcez and Luis C. Lamb. “Neurosymbolic AI: The 3rd Wave”. In: *arXiv:2012.05876 [cs]* (Dec. 16, 2020). arXiv: 2012.05876. URL: http://arxiv.org/abs/2012.05876 (visited on 02/23/2022).

[35] Peter Gärdenfors. *Conceptual spaces: The geometry of thought*. Conceptual spaces: The geometry of thought. Cambridge, MA, US: The MIT Press, 2000. x, 307. ISBN: 978-0-262-07199-4.

[36] Kurt Gödel. “Zum intuitionistischen aussagenkalkül”. In: *Anzeiger der Akademie der Wissenschaften in Wien* 69 (1932).

[37] Michael Goodale. "Do contextual word embeddings represent richly subsective adjectives more diversely than intersective adjectives?" Bridges and Gaps Workshop ESSLLI 2022. Galway, Ireland, Aug. 8, 2022.

[38] Nelson Goodman. "Seven Strictures on Similarity". In: *Problems and Projects*. Bobs-Merril, 1972.

[39] Janek Guerrini. "Similarity statements, privative adjectives, and generics: a unified view". Qualifying Paper. Ecole Normale Superieure, Aug. 2021. URL: https://lingbuzz.net/lingbuzz/006221 (visited on 11/28/2021).

[40] Janek Guerrini. "The Link between Misinterpretation, Intentionality, and Mental Agency in the Natural Language Interpretation of "Fake"". In: *Rivista Italiana di Filosofia Analitica Junior* 9.2 (Dec. 31, 2018). Number: 2, pp. 181–192. ISSN: 2037-4445. DOI: 10.13130/2037-4445/11095. URL: https://riviste.unimi.it/index.php/rifanalitica/article/view/11095 (visited on 06/06/2022).

[41] Janek Guerrini and Salvador Mascarenhas. "Shifting centers: toward a unified view of grammatical and contingent privative modification". Università di Siena, Sept. 26, 2019.

[42] Petr Hájek. *Metamathematics of fuzzy logic*. Trends in logic v. 4. Dordrecht ; Boston: Kluwer, 1998. 297 pp. ISBN: 978-0-7923-5238-9.

[43] James A. Hampton. "Inheritance of attributes in natural concept conjunctions". In: *Memory & Cognition* 15.1 (Jan. 1, 1987), pp. 55–71. ISSN: 1532-5946. DOI: 10.3758/BF03197712. URL: https://doi.org/10.3758/BF03197712 (visited on 06/01/2022).

[44] Philip Johnson-Laird. "Formal Semantics and the Psychology of Meaning". In: *Processes, Beliefs, and Questions: Essays on Formal Semantics of Natural Language and Natural Language Processing*. Ed. by Stanley Peters and Esa Saarinen. Synthese Language Library. Dordrecht: Springer Netherlands, 1982, pp. 1–68. ISBN: 978-94-015-7668-0. DOI: 10.1007/978-94-015-7668-0_1. URL: https://doi.org/10.1007/978-94-015-7668-0_1 (visited on 06/04/2022).

[45] Mark Johnston and Sarah-Jane Leslie. "Cognitive Psychology and the Metaphysics of Meaning". In: *Metaphysics and Cognitive Science*. Oxford University Press, June 27, 2019, pp. 183–205. ISBN: 978-0-19-063967-9 978-0-19-063970-9. DOI: 10.1093/oso/9780190639679.003.0008. URL: https://oxford.universitypressscholarship.com/view/10.1093/oso/9780190639679.001.0001/oso-9780190639679-chapter-8 (visited on 11/28/2021).

[46] Hans Kamp. “Two theories about adjectives”. In: *Formal Semantics of Natural Language*. Ed. by Edward L. Keenan. Cambridge: Cambridge University Press, 1975, pp. 123–155. ISBN: 978-0-521-11111-9. DOI: 10.1017/CBO9780511897696.011. URL: https://www.cambridge.org/core/books/formal-semantics-of-natural-language/two-theories-about-adjectives/DA68C05CF9AE0F0FC98CC73C79F844C0 (visited on 05/16/2022).

[47] Hans Kamp and Barbara Partee. “Prototype theory and compositionality”. In: *Cognition* 57.2 (Nov. 1995), pp. 129–191. ISSN: 00100277. DOI: 10.1016/0010-0277(94)00659-9. URL: https://linkinghub.elsevier.com/retrieve/pii/0010027794006599 (visited on 11/28/2021).

[48] Frank C. Keil. *Concepts, kinds, and cognitive development*. Concepts, kinds, and cognitive development. Cambridge, MA, US: The MIT Press, 1989. xv, 328. ISBN: 978-0-262-11131-7.

[49] Christopher Kennedy. “Adjectives”. In: *The Routledge companion to philosophy of language*. Ed. by Gillian Russell and Delia Graff Fara. First published in paperback. Routledge philosophy companions. New York London: Routledge, 2015. ISBN: 978-1-138-77618-0 978-0-415-99310-4.

[50] Christopher Kennedy. “Vagueness and grammar: the semantics of relative and absolute gradable adjectives”. In: *Linguistics and Philosophy* 30.1 (Feb. 1, 2007), pp. 1–45. ISSN: 1573-0549. DOI: 10.1007/s10988-006-9008-0. URL: https://doi.org/10.1007/s10988-006-9008-0 (visited on 04/11/2022).

[51] Christopher Kennedy and Louise McNally. “Scale Structure, Degree Modification, and the Semantics of Gradable Predicates”. In: *Language* 81.2 (2005), pp. 345–381. ISSN: 1535-0665. DOI: 10.1353/lan.2005.0071. URL: http://muse.jhu.edu/content/crossref/journals/language/v081/81.2kennedy.pdf (visited on 05/15/2022).

[52] Sangeet Khemlani and P. N. Johnson-Laird. “Disjunctive illusory inferences and how to eliminate them”. In: *Memory & Cognition* 37.5 (July 1, 2009), pp. 615–623. ISSN: 1532-5946. DOI: 10.3758/MC.37.5.615. URL: https://doi.org/10.3758/MC.37.5.615 (visited on 05/03/2022).

[53] Durk P Kingma and Prafulla Dhariwal. “Glow: Generative Flow with Invertible 1x1 Convolutions”. In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc., 2018. URL: https://papers.nips.cc/paper/2018/hash/d139db6a236200b21cc7f752979132d0-Abstract.html (visited on 05/25/2022).

[54] Arnold Kochari, Robert van Rooij, and Katrin Schulz. "Generics and Alternatives". In: *Frontiers in Psychology* 11 (2020), p. 1274. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2020.01274. URL: https://www.frontiersin.org/article/10.3389/fpsyg.2020.01274 (visited on 01/04/2022).

[55] Philipp Koralus and Salvador Mascarenhas. "The Erotetic Theory of Reasoning". In: *Philosophical Perspectives* 27.1 (Dec. 2013), pp. 312–365. ISSN: 15208583. DOI: 10.1111/phpe.12029. URL: https://onlinelibrary.wiley.com/doi/10.1111/phpe.12029 (visited on 03/20/2022).

[56] Manfred Krifka and Claudia Gerstner. "An outline of genericity". In: Seminar für natürlich-sprachliche Systeme der Universität Tübingen, 1987.

[57] Manfred Krifka, Francis Jeffry Pelletier, et al. "Genericity: An Introduction". In: *The Generic Book*. Ed. by Greg N. Carlson and Francis Jeffry Pelletier. University of Chicago Press, 1995, pp. 1–124.

[58] Saul A. Kripke. *Naming and necessity*. Cambridge, Mass: Harvard University Press, 1980. 172 pp. ISBN: 978-0-674-59845-4.

[59] Saul A. Kripke. "Speaker's Reference and Semantic Reference". In: *Studies in the Philosophy of Language*. Ed. by Peter A. French, Theodore E. Uehling Jr, and Howard K. Wettstein. University of Minnesota Press, 1977, pp. 255–296.

[60] J.M. Lawler. *Studies in English Generics*. University of Michigan, 1974. URL: https://books.google.fr/books?id=RAnTxAEACAAJ.

[61] John M. Lee. *Riemannian Manifolds*. Vol. 176. Graduate Texts in Mathematics. New York, NY: Springer New York, 1997. ISBN: 978-0-387-98322-6 978-0-387-22726-9. DOI: 10.1007/b98852. URL: http://link.springer.com/10.1007/b98852 (visited on 06/02/2022).

[62] Sarah-Jane Leslie. "Generics and the Structure of the Mind". In: *Philosophical Perspectives* 21.1 (2007), pp. 375–403. ISSN: 1520-8583. DOI: 10.1111/j.1520-8583.2007.00138.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1520-8583.2007.00138.x (visited on 11/24/2021).

[63] Sarah-Jane Leslie, Sangeet Khemlani, and Sam Glucksberg. "Do all ducks lay eggs? The generic overgeneralization effect". In: *Journal of Memory and Language* 65.1 (July 1, 2011), pp. 15–31. ISSN: 0749-596X. DOI: 10.1016/j.jml.2010.12.005. URL: https://www.sciencedirect.com/science/article/pii/S0749596X10001154 (visited on 01/27/2022).

[64] Sarah-Jane Leslie and Adam Lerner. “Generic Generalizations”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2016. Metaphysics Research Lab, Stanford University, 2016. URL: https://plato.stanford.edu/archives/win2016/entries/generics/ (visited on 06/06/2022).

[65] David K. Lewis. “Adverbs of Quantification”. In: *Formal Semantics of Natural Language*. Ed. by Edward L. Keenan. Cambridge University Press, 1975, pp. 3–15.

[66] David Liebesman. “Simple Generics”. In: *Noûs* 45.3 (2011), pp. 409–442. ISSN: 0029-4624. URL: https://www.jstor.org/stable/41330866 (visited on 01/26/2022).

[67] Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. “Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation”. In: *Journal of Memory and Language* 92 (Feb. 1, 2017), pp. 57–78. ISSN: 0749-596X. DOI: 10.1016/j.jml.2016.04.001. URL: https://www.sciencedirect.com/science/article/pii/S0749596X16300079 (visited on 06/04/2022).

[68] Gary Marcus. *Deep Learning: A Critical Appraisal*. arXiv:1801.00631. type: article. arXiv, Jan. 2, 2018. DOI: 10.48550/arXiv.1801.00631. arXiv: 1801.00631[cs,stat]. URL: http://arxiv.org/abs/1801.00631 (visited on 06/06/2022).

[69] Joshua Martin. “Compositionality in Privative Adjectives: Extending Dual Content Semantics”. In: *At the Intersection of Language, Logic, and Information*. Ed. by Jennifer Sikos and Eric Pacuit. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2019, pp. 93–107. ISBN: 978-3-662-59620-3. DOI: 10.1007/978-3-662-59620-3_6.

[70] Salvador Mascarenhas. “Formal Semantics and the Psychology of Reasoning: Building new bridges and investigating interactions”. PhD thesis. NYU, Sept. 2014. URL: https://ling.auf.net/lingbuzz/002213 (visited on 06/04/2022).

[71] Lisa Matthewson. “Quantification and the Nature of Crosslinguistic Variation”. In: *Natural Language Semantics* 9.2 (June 1, 2001), pp. 145–189. ISSN: 1572-865X. DOI: 10.1023/A:1012492911285. URL: https://doi.org/10.1023/A:1012492911285 (visited on 01/27/2022).

[72] Tom McCoy, Ellie Pavlick, and Tal Linzen. “Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL 2019. Association for Computational Linguistics, July 2019, pp. 3428–3448. DOI: 10.18653/v1/P19-1334. URL: https://aclanthology.org/P19-1334 (visited on 10/17/2021).

[73] Louise McNally. “Modification”. In: *The Cambridge Handbook of Formal Semantics*. Ed. by Maria Aloni and Paul Dekker. Cambridge Handbooks in Language and Linguistics. Cambridge: Cambridge University Press, 2016, pp. 442–464. ISBN: 978-1-139-23615-7. DOI: 10.1017/CBO9781139236157.016. URL: https://www.cambridge.org/core/books/cambridge-handbook-of-formal-semantics/modification/326E443196570780733CC319B0A68B73 (visited on 05/17/2022).

[74] Tomas Mikolov et al. “Efficient Estimation of Word Representations in Vector Space”. In: *arXiv:1301.3781 [cs]* (Sept. 6, 2013). arXiv: 1301.3781. URL: http://arxiv.org/abs/1301.3781 (visited on 03/27/2021).

[75] Richard Montague. “English as a Formal Language”. In: *Linguaggi nella societa e nella tecnica*. Ed. by Bruno Visentini. Edizioni di Communita, 1970, pp. 188–221.

[76] Marcin Morzycki. *Modification*. Key Topics in Semantics and Pragmatics. Cambridge: Cambridge University Press, 2015. ISBN: 978-1-107-00975-2. DOI: 10.1017/CBO9780511842184. URL: https://www.cambridge.org/core/books/modification/ECF7ED25FD3A537840AAE6BF30B42771 (visited on 04/05/2022).

[77] Maximillian Nickel and Douwe Kiela. “Poincaré Embeddings for Learning Hierarchical Representations”. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. URL: https://papers.nips.cc/paper/2017/hash/59dfa2df42d9e3d41f5b02bfc32229dd-Abstract.html (visited on 05/11/2022).

[78] Håkan Nilsson, Peter Juslin, and Henrik Olsson. “Exemplars in the mist: The cognitive substrate of the representativeness heuristic”. In: *Scandinavian Journal of Psychology* 49.3 (2008), pp. 201–212. ISSN: 1467-9450. DOI: 10.1111/j.1467-9450.2008.00646.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9450.2008.00646.x (visited on 11/23/2021).

[79] Robert M. Nosofsky. “Exemplar-based accounts of relations between classification, recognition, and typicality”. In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14.4 (1988), pp. 700–708. ISSN: 1939-1285. DOI: 10.1037/0278-7393.14.4.700.

[80] Daniel N. Osherson and Edward E. Smith. “On the adequacy of prototype theory as a theory of concepts”. In: *Cognition* 9.1 (1981), pp. 35–58. ISSN: 1873-7838. DOI: 10.1016/0010-0277(81)90013-5.

[81] T. Parsons. “Events in the Semantics of English: A Study in Subatomic Semantics”. In: *undefined* (1990). URL: https://www.semanticscholar.org/paper/Events-in-the-Semantics-of-English%3A-A-Study-in-Parsons/8ff703240303808d2d54b0d7723550820d17a7ca (visited on 01/17/2022).

[82] Barbara Partee. “Privative Adjectives: Subsective Plus Coercion”. In: *Presuppositions and Discourse: Essays Offered to Hans Kamp* (Jan. 1, 2010), pp. 273–285. DOI: 10.1163/9789004253162_011. URL: https://brill.com/view/book/edcoll/9789004253162/B9789004253162-s011.xml (visited on 11/28/2021).

[83] Matthew E. Peters et al. “Semi-supervised sequence tagging with bidirectional language models”. In: *arXiv:1705.00108 [cs]* (Apr. 28, 2017). arXiv: 1705.00108. URL: http://arxiv.org/abs/1705.00108 (visited on 10/17/2021).

[84] Paul M. Pietroski. *Conjoining Meanings: Semantics Without Truth Values*. Context & Content. Oxford, New York: Oxford University Press, Apr. 26, 2018. 404 pp. ISBN: 978-0-19-881272-2.

[85] James Pustejovsky. *The Generative Lexicon*. Language, Speech, and Communication. Cambridge, MA, USA: A Bradford Book, Oct. 23, 1995. 312 pp. ISBN: 978-0-262-16158-9.

[86] Hilary Putnam. “Meaning and Reference”. In: *The Journal of Philosophy* 70.19 (1973), pp. 699–711. ISSN: 0022-362X. DOI: 10.2307/2025079. URL: https://www.jstor.org/stable/2025079 (visited on 11/30/2021).

[87] M. M. G. Ricci and T. Levi-Civita. “Méthodes de calcul différentiel absolu et leurs applications”. In: *Mathematische Annalen* 54.1 (Mar. 1, 1900), pp. 125–201. ISSN: 1432-1807. DOI: 10.1007/BF01454201. URL: https://doi.org/10.1007/BF01454201 (visited on 05/25/2022).

[88] Iris van Rooij et al. “Rational analysis, intractability, and the prospects of ‘as if’-explanations”. In: *Synthese* 195.2 (Feb. 2018), pp. 491–510. ISSN: 0039-7857, 1573-0964. DOI: 10.1007/s11229-014-0532-0. URL: http://link.springer.com/10.1007/s11229-014-0532-0 (visited on 05/25/2022).

[89] Robert van Rooij and Katrin Schulz. “Generics and typicality: a bounded rationality approach”. In: *Linguistics and Philosophy* 43.1 (Feb. 2020), pp. 83–117. ISSN: 0165-0157, 1573-0549. DOI: 10.1007/s10988-019-09265-8. URL: http://link.springer.com/10.1007/s10988-019-09265-8 (visited on 01/04/2022).

[90] Eleanor Rosch. “Cognitive Representations of Semantic Categories”. In: *Journal of Experimental Psychology: General* 104.3 (1975), pp. 192–233. DOI: 10.1037/0096-3445.104.3.192.

[91] Eleanor Rosch et al. “Basic objects in natural categories”. In: *Cognitive Psychology* 8.3 (July 1, 1976), pp. 382–439. ISSN: 0010-0285. DOI: 10.1016/0010-0285(76)90013-X. URL: https://www.sciencedirect.com/science/article/pii/001002857690013X (visited on 05/11/2022).

[92] Bertrand Russell. “On Denoting”. In: *Mind* 14.56 (1905). Publisher: [Oxford University Press, Mind Association], pp. 479–493. ISSN: 0026-4423. URL: https://www.jstor.org/stable/2248381 (visited on 06/04/2022).

[93] Mathias Sablé-Meyer and Salvador Mascarenhas. “Indirect Illusory Inferences From Disjunction: A New Bridge Between Deductive Inference and Representativeness”. In: *Review of Philosophy and Psychology* (2021), pp. 1–26. DOI: 10.1007/s13164-021-00543-8.

[94] Uli Sauerland. “Vagueness in Language: The Case Against Fuzzy Logic Revisited”. In: Jan. 1, 2011, pp. 185–198.

[95] Shreya Saxena and John P Cunningham. “Towards the neural population doctrine”. In: *Current Opinion in Neurobiology* 55 (Apr. 2019), pp. 103–111. ISSN: 09594388. DOI: 10.1016/j.conb.2019.02.002. URL: https://linkinghub.elsevier.com/retrieve/pii/S0959438818300990 (visited on 05/23/2022).

[96] Roger N. Shepard. “The analysis of proximities: Multidimensional scaling with an unknown distance function. I.” In: *Psychometrika* 27.2 (June 1, 1962), pp. 125–140. ISSN: 1860-0980. DOI: 10.1007/BF02289630. URL: https://doi.org/10.1007/BF02289630 (visited on 05/11/2022).

[97] Roger N. Shepard. “Toward a Universal Law of Generalization for Psychological Science”. In: *Science* 237.4820 (Sept. 11, 1987), pp. 1317–1323. DOI: 10.1126/science.3629243. URL: https://www.science.org/doi/10.1126/science.3629243 (visited on 05/11/2022).

[98] Edward E. Smith and Daniel N. Osherson. "Conceptual combination with prototype concepts". In: *Cognitive Science* 8.4 (Oct. 1, 1984), pp. 337–361. ISSN: 0364-0213. DOI: 10.1016/S0364-0213(84)80006-3. URL: https://www.sciencedirect.com/science/article/pii/S0364021384800063 (visited on 11/23/2021).

[99] Ryota Suzuki, Ryusuke Takahama, and Shun Onoda. "Hyperbolic Disk Embeddings for Directed Acyclic Graphs". In: *arXiv:1902.04335 [cs, stat]* (May 15, 2019). arXiv: 1902.04335. URL: http://arxiv.org/abs/1902.04335 (visited on 11/23/2021).

[100] Anna Szabolcsi. *Quantification*. Cambridge: Cambridge University Press, 2010. ISBN: 978-0-511-78168-1. DOI: 10.1017/CBO9780511781681. URL: http://ebooks.cambridge.org/ref/id/CBO9780511781681 (visited on 06/01/2022).

[101] Amos Tversky. "Features of similarity." In: *Psychological Review* 84.4 (1977), pp. 327–352. ISSN: 0033-295X. DOI: 10.1037/0033-295X.84.4.327. URL: http://content.apa.org/journals/rev/84/4/327 (visited on 05/11/2022).

[102] Amos Tversky and Daniel Kahneman. "Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment." In: *Psychological Review* 90.4 (1983), pp. 293–315. ISSN: 0033-295X. DOI: 10.1037/0033-295X.90.4.293. URL: http://content.apa.org/journals/rev/90/4/293 (visited on 11/30/2021).

[103] Amos Tversky and Daniel Kahneman. "Judgment under Uncertainty: Heuristics and Biases". In: *Science* 185.4157 (1974), pp. 1124–1131. ISSN: 0036-8075. URL: https://www.jstor.org/stable/1738360 (visited on 11/27/2021).

[104] Ivan Vendrov et al. "Order-Embeddings of Images and Language". In: *arXiv:1511.06361 [cs]* (Mar. 1, 2016). arXiv: 1511.06361. URL: http://arxiv.org/abs/1511.06361 (visited on 03/30/2022).

[105] Clare R. Walsh and P. N. Johnson-Laird. "Co-reference and reasoning". In: *Memory & Cognition* 32.1 (Jan. 2004), pp. 96–106. ISSN: 0090-502X. DOI: 10.3758/bf03195823.

[106] Fei Xu and Joshua B. Tenenbaum. "Word learning as Bayesian inference." In: *Psychological Review* 114.2 (2007), pp. 245–272. ISSN: 1939-1471, 0033-295X. DOI: 10.1037/0033-295X.114.2.245. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.114.2.245 (visited on 05/11/2022).

[107] Ronald R. Yager and Alexander Rybalov. "Uninorm aggregation operators". In: *Fuzzy Sets and Systems*. Fuzzy Modeling 80.1 (May 27, 1996), pp. 111–120. ISSN: 0165-0114. DOI: 10.1016/0165-0114(95)00133-6. URL: https://www.sciencedirect.com/science/article/pii/0165011495001336 (visited on 05/22/2022).

[108] L.A. Zadeh. "Fuzzy sets". In: *Information and Control* 8.3 (June 1965), pp. 338–353. ISSN: 00199958. DOI: 10.1016/S0019-9958(65)90241-X. URL: https://linkinghub.elsevier.com/retrieve/pii/S001999586590241X (visited on 05/03/2022).

Appendix A

Pre-Registration

Concepts in Hyperbolic Space

A Model of Lexical Meaning for Model-Theoretic Semantics

A.1 Administration

Possible reader from the conseil pédagogique: Emmanuel Chemla

Possible external reader: Gemma Boleda

A.2 Introduction

The core primitives which are used in model-theoretic semantics are the set and the individual, e.g. l_1 is a lion and so it is in the set of all lions, $\{l_1, l_2, \ldots\}$. This set extensional view of concepts is a very useful abstraction for formal work on semantic composition. For example, we can say that "a cat chases a mouse" has the logical form, $\exists x\, (cat(x) \wedge \exists y\, (mouse(y) \wedge chased(x,y)))$ where each predicate is true or false depending on if the given individual is in the set of that predicate.

The set extensional view of lexical meaning is a very useful abstraction, but it has many problems. First, it is not very cognitive plausible, especially if one takes an externalist view of semantics, without subscribing to a kind of Fodorian nativism (Fodor 1981). Furthermore, within model-theoretic semantics, there are problems in semantics where the notion of meaning on a *lexical* level keeps popping up where sets are not enough, and there needs to be a notion of typicality or similarity which cannot be accounted for using sets.

The goal of my research is to propose a new model which using insights from psychology, natural language processing and semantics to produce a revision to traditional semantic theory which revises the basic primitives used. The goal is to provide a new *formal* model of concepts which will be able to account for some of these shortcomings using theoretical and computational models from psychology and natural language processing.

Core question: What semantic and reasoning problems can be accounted for by describing the primitives used in traditional semantics (e.g. sets and individuals) as regions and points in a high-dimensional conceptual space?

A.3 Methods

A.3.1 Conceptual sources

Since the work is largely about consolidating disparate accounts of lexical concepts in a unified theory, there will necessarily be a broad set of sources to build from.

Psychology

A large portion will be drawn from psychology, in particular prototype theory (Rosch 1975), exemplar theory (Nosofsky 1988), perceptual symbol systems (Barsalou 1999) and conceptual spaces (Gärdenfors 2000). One thing that links these different approaches is that they are atheoretical, in other words, they do not assume that people construct logical theories about words but that the definition is more hazy. There are certainly somethings that people *must* build theories about, such as numbers. These kinds of words are not the main focus of the research project, but it is still necessary to be informed by theory-theory approaches such as Carey 2009.

NLP

Some important tools used are things such as contextual word embeddings (Peters et al. 2017). Another important concept is grounded distributional semantics (Baroni 2016) where word meanings are made distributionally but that distribution is extended to other forms of data (e.g. images, audio) which would better represent the rich multi-sensory world that humans typically acquire language in.

Finally, the core structure of the model is adapted from Suzuki, Takahama, and Onoda 2019 because it provides a embedding of a partial order in a continuous space.

Semantics

Semantic work that will be involved are previous attempts to introduce qualitative features to a formal semantics (Pustejovsky 1995; Del Pinal 2015) or that try and critique such approaches (Guerrini 2021) and finally works which aim to understand semantics on an internalist approach (Pietroski 2018) rather than externalist. For the specific applications of the approach such as generics, it will be important to contrast works which *do* assume typicality-based explanations (Leslie 2007) and those which prefer to continue to use raw set extensional accounts (Kochari, R. v. Rooij, and Schulz 2020; Cohen 1996).

A.3.2 A formal model of concepts

The formal model proposed has several desiderata to which the model should hold:

The feature space shouldn't necessarily be interpretable. The only innate and immediately interpretable features should be perceptual data or innate feature recognizers (Carey 2009), we do not assume a featural representation of lexical concepts such as "mane".

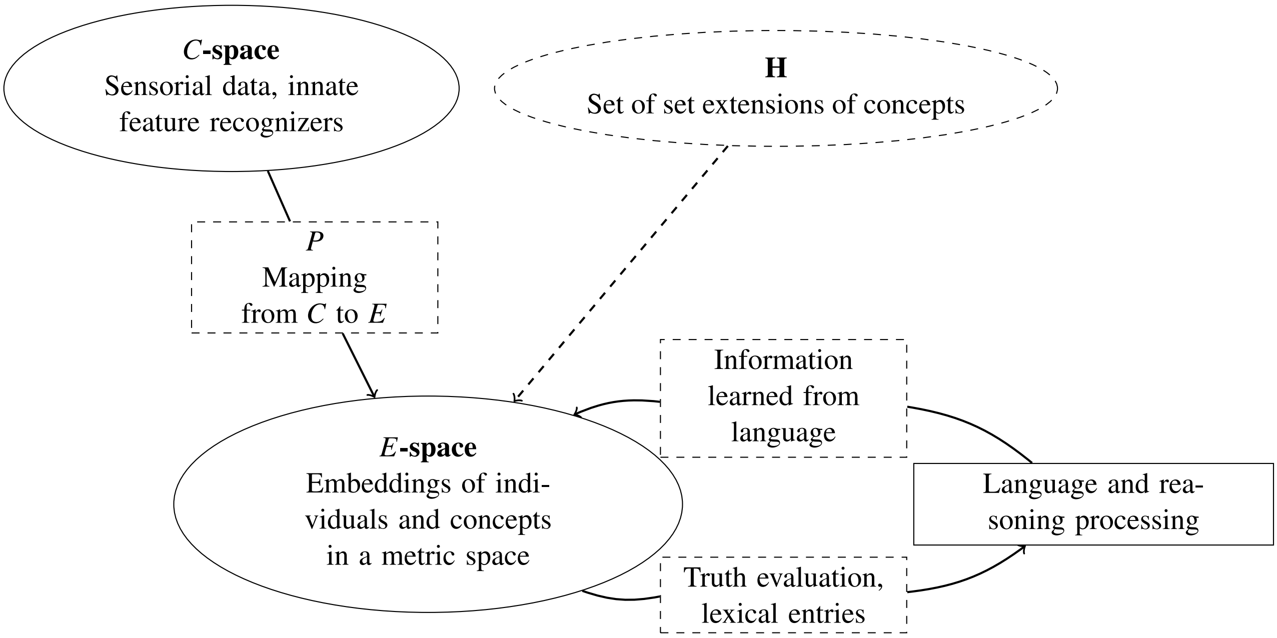
Judgements about membership to a category should be all or nothing. While people can debate whether something belongs to a category or not; they are debating a fact, not the degree to which it is in a category. Contrary to prototype theory or exemplar theory, belonging to a category is *categorical* and should be modeled as such.

Some individuals should be considered more or less typical within their category. Closely related to the previous statement, some things may belong to a category yet be atypical. For example, while penguins are birds, they are clearly not typical birds.

Typicality should be related to similarity and implicated in reasoning and linguistic judgements This typicality metric must be related to how people judge the similarity between individuals as well (Guerrini 2021). For example, the conjunction fallacy (Tversky and Kahneman 1974) is often explained using a representativeness heuristic or a typicality metric.

We should still be able to use it in semantic theories. There should be a way to extract set extensions (or something analogous) in order to evaluate truth-values of statements.

A.3.3 Outline of model



C-space
Sensorial data, innate feature recognizers
H
Set of set extensions of concepts
P
Mapping from *C* to *E*
Information learned from language
E-space
Embeddings of individuals and concepts in a metric space
Language and reasoning processing
Truth evaluation, lexical entries

Figure A.1: Flowchart of the different spaces and their relationships used in this approach.

Featural space

Given a something, x, it can be characterised as a vector, $\vec{x}$ in C where C is a vector space in $\mathbb{R}^n$. The dimensions of C simply correspond to the various neurons that are integral to sensorial processing whether high or low-level, e.g. output of visual cortex, or low-level, e.g. neurons in the retina. $\vec{x}$ would then be a vector which represents the activation of various neurons upon seeing x.

A.3.4 Set extensions

We define a set of sets, H, where each set in H is a set extension for a given concept, e.g. `Cat`, `Dog`, etc. We equip H with a partial order, $\preceq$. This ordering consists of whether a concept is a hypernym of the other, e.g. `Cat`$\preceq$`Animal`.

A.3.5 Conceptual space

In order to connect the C-space and H, we need a way to express concepts in a continuous space. Given a concept $y \in H$, it can also be characterised as a point, $\vec{y}$ which lies in a space, E. In order to get a partial ordering in E, we use Hyperbolic Disk Embeddings (Suzuki, Takahama, and Onoda 2019).

For each concept y in H, there is associated point, $\vec{y}$ in E and a positive real-valued radius, r_y. These define a ball, $D(\vec{y}, r_y)$ for each concept. The core idea is that given $x \in H$ and $y \in H$, $x \preceq y$ if and only if $D(\vec{x}, r_x) \subseteq D(\vec{y}, r_y)$. In other words, like a Euler diagram, concept x is a hypernym of concept y if and only if the area of y lies entirely within the area of x.

Since E implicitly implements H, we don't need H to ever be represented directly.

A.3.6 Individuals

We also need a way to define individuals, i.e. that particular lion, or the things which belong to the sets of H. We define these as conceptual balls where the radius is zero, since individuals are the minima of the partial ordering of H. So, a particular lion will be a point which lies within the radius of the lion concept ball.

A.3.7 Conceptual map

We now define a function, P which maps from C to E. In other words, given a specific percept or the features of a specific individual, we find an associated vector in E.

Similarity

To determine how similar two concepts are, we can simply take the metric distance of the two concepts. This also works for individual percepts, e.g. comparing two different lions.

Prototypicality

A similar procedure can be used to see how prototypical an individual is of a category.

$$\text{Prototypicality}(x, c) = \max\left(\frac{r_c - d(P(\vec{x}), \vec{c})}{r_c}, 0\right) \tag{A.1}$$

Here, prototypicality is simply the distance between an individual and the centre of a concept scaled with respect to the concept's radius. This has the property where the prototypicality of an individual is one if it maps perfectly to the centre of a concept. As the individual gets closer and closer to the edge of the concept's ball (and becomes more atypical), it approaches 0, until it reaches the very edge of the concept, where its representativeness is then 0. After that, the representativeness is 0, since it is not a member of that concept.

A.3.8 Theoretical problems to account for

There are several linguistic problems which may become more clear by enriching the basic semantic primitives:

1. Generic sentences certainly don't correspond to a straightforward quantification over sets. They may represent cognitive "defaults" (Leslie 2007) or probability judgements over unbounded sets (Kochari, R. v. Rooij, and Schulz 2020) or the most "typical" possible worlds (Asher and Morreau 1995).

   (a) Cats have tails (well, some don't but most do).

   (b) Ducks lay eggs (less than 50% do!)

   (c) Mosquitoes carry the West Nile Virus (very few do!)

2. Intensional adjectives somehow negate the set they belong to. A fake gun is *like* a gun in some capacity but vitally not a gun.

3. Noun-noun composition (stone-lions are statues of lions, rather than something that is both a stone and a lion)

4. Even simple conditionals can rely on notions of "typical" which are not defined in a rigorous way.

   (a) If I strike a match, it will light. ✓

   (b) If I strike a match and I'm underwater, it will light. ✗

The main goal of this theoretical work is to see if adopting the described model can give convincing accounts of these linguistic phenomena in a mathematically precise way.

The other main goal is to demonstrate that it is interoperable traditional model-theoretic semantics and so it will be necessary to provide definitions of the tools of model theoretic semantics (e.g. quantification). Some of the most important ones that I will attempt to outline are the following, although this is contingent on the success of the more simple goals such as quantification.

- Scalar adjectives. These may be best modeled as a function which takes a point in E-space and returns a real number.
- n-ary predicates.[1]
- Event semantics (this may involve two different E-spaces, one for events and one for individuals)
- Modal logic and possible worlds.

A.3.9 Novelty

The specific novelty of this model is that it accounts for typicality effects in language while preserving the rich, compositional framework provided by using a model-theoretic approach to semantics. Furthermore, it links lexical concepts to external perception in a motivated way and so doesn't require sub-lexical features which are interpretable. Crucially, it differs from most theories of concepts from psychology because it provides definitions for the logical computations necessary for model-theoretic semantics.

A.3.10 Interpretation

This research project could be extremely broad, stressing the importance of defining a precise goal. A minimally successful proof of concept could be to provide an explanation of generic sentences and a way to express monadic first order logic. Vitally, this account of generics should not fall prey to the typical problems with set-quantification accounts of generics outlined by Leslie 2007. More ambitious goals have been described in the previous sections.

[1] One interesting approach could be to look at Paul Pietroski who argues that most (or all) lexical concepts are monadic, and that n-ary predicates are created by composing different basic n-ary predicates(e.g. thematic roles) along with the lexical terms (Pietroski 2018) in an approach similar to the neo-Davidsonian view of event semantics (Parsons 1990; Krifka and Gerstner 1987). For example, "John loves Mary" could mean $\exists e(love(e) \land agent(e, j) \land patient(e, m))$.