Language Games With Negations for Vague Categories

Henrietta Eyre,
Black Swan Data,
11 York Road,
Tower Building, 6th Floor,
London SE1 7NX, UK
email: henrietta.eyre@blackswan.com

Jonathan Lawry,
Department of Engineering Mathematics,
University of Bristol,
Bristol BS8 1TR, UK
email: j.lawry@bris.ac.uk

Abstract

The guessing game is a language game which models the communication protocols between two agents who aim to match category labels with objects encountered in a simulated or real environment. Here we present a new representation framework for guessing games, in which category definitions explicitly incorporate semantic uncertainty and typicality. More specifically, we propose a conceptual model based on prototype and random set theory, in which categories are defined within a metric conceptual space. We argue that this conceptual framework is both expressive and naturally generates robust assertion and concept updating models. In particular, we define both assertion and updating rules for guessing games involving a mixture of labels and negated labels. Finally, the results of language game simulations are presented, where a multi-agent system evolves through pairwise language games incorporating an assertion and an updating algorithm. Our results suggest that, within this framework, a mixture of both positive and negative assertions may be required in order for agent interpretations to converge, whilst retaining sufficiently discriminatory categories for effective communication.

Keywords Language games, guessing game, semantic uncertainty, prototype theory, multi-agent systems, conceptual spaces.

1 Introduction

Language games offer a model under which artificial systems may take account of the evolutionary nature of language learning [51]. In a population of agents playing language games, the syntactic and semantic structures may be individual to each agent, but groups of agents then cooperate to evolve common structures and methods of communication.
The underlying hypothesis then is that an evolutionary approach to determining a shared language will result in communications which are more efficient than if a designer attempted to preprogram agents with a fixed syntax and semantics.

In this paper we present an investigation of the guessing game. We will investigate the emergence of categories which are inherently vague. Here we associate vagueness with blurred boundaries and adopt a probabilistic approach consistent with those proposed by Hisdal [26], Edgington [15], Lawry [31], Bennett [5], Lassiter [30], and Goodman and Lassiter [22]. From this perspective vagueness is seen as the result of uncertainty about the correct definition of categories which naturally arises as an integral part of the concept learning process.

Spranger and Pauw have presented extensive investigations into vague categories in language games [48]. Our conceptual model also describes vague categories with uncertain category boundaries, but also allows for categories to be situated in a conceptual space so that there may be a region of the conceptual space whose conceptual definition is not covered by any category. Under this conceptual model we present a new investigation to the literature, where agents may describe the focus of a language game using negated labels as well as basic label descriptions. We argue that it is both realistic and useful to adopt a this richer assertion set.

An outline of the paper is as follows. Section 2 gives an overview of the language games literature, section 3 discusses cognitive models, Gärdenfors’ conceptual space model and the random set and prototype model of concepts which we use as the representation framework for our language games. Section 4 gives a description of our language game model, including assertion and updating algorithms. Section 5 describes the metrics which quantify the behaviour of the language game system during the experimental simulations. Section 6 presents the results from our simulation studies, and finally section 7 gives some discussion and conclusions.

2 Language Games

In this paper we base our simulated language systems on the notion of a language game in the sense of Steels [50]. In this context, a language game describes a full communicative interaction between two agents, described by Steels as the “full semiotic cycle” [50]. One agent, acting as a speaker, will formulate a linguistic utterance to be asserted to the interacting agent (the listener) given the agent’s conceptual model, a goal and the constraints placed upon this goal (for example by the environment). The listener must then recognise the assertion they receive, interpret its meaning in relation to their conceptual model and update this model to satisfy any constraints implied by the assertion.

Two types of multi-agent language interaction which have attracted significant attention in the literature are the Naming Game and the Category Game. The Naming Game was first introduced by Steels [53, 49] to explore the role of self-organisation in language. A population agree on a given convention [34] (for example the name assigned to a particular object) by relying only on local interactions. Each agent is equipped with a discrete set of labels for the given object and at each time step two agents are selected to interact, one as the speaker and one as the listener. A linguistic convention evolves through these
interactions where a speaker chooses a word in their dictionary so as to describe the object to the listener. If both agents have this word in their dictionary then all other words are dropped from their dictionaries, resulting in a maximal reduction in redundant words for the given pair of agents. Alternatively, if the speaker is transmitting an unknown word to the hearer, they then add it to their dictionary. Such a system will converge on a unified vocabulary, as shown both analytically and through simulation in [56]. The naming game has attracted attention from the statistical physics community, who have investigated it under different assumptions about population structure and communication rules. In particular, small world networks [10, 35], static [11] and dynamic [37] complex networks, have all been studied in this context resulting in a number of convergence result similar to those found for fully connected networks [49].

The guessing game does not focus on one object of interest but rather the set of all objects within a given attribute space. In this way it aims to investigate category formation and mechanisms for categorisation by monitoring the evolution of a system where agents have a set of basic communication rules and (initially) an empty set of concepts. Through the course of the game concepts emerge and evolve through basic communication between agents. This was first proposed by Steels [54] who demonstrated that there are models according to which a population evolves so that the concept definitions of different agents become sufficiently similar so as to allow for effective communication. A typical guessing game involves a population of players, each of whose concepts are modelled as partitions of the unit interval. The dynamics of agent interaction then results in the manipulation of these discrete intervals by linking perceptual categories to words presented to interacting agents, and where new categories may be introduced or the width of existing categories extended.

The guessing game is often performed on colour categories as modelled in a continuous space (e.g. CIE Lab space for a conceptually accurate model) making them easy to implement in a simulated language game [54]. Such investigations can make use of the World Color Survey (WCS) [8], a catalog of language data from dozens of languages allowing comparison between the emergent patterns in simulated guessing games and the patterns observed in the WCS.

Belpaeme and Bleys showed in [4] that colour terms may emerge in a system with cultural input. Their results show remarkable agreement with the WCS. In addition, [14] demonstrates the robustness of the adaptive language system (for colour terms) under random noise. By applying a Bayesian inference procedure, the system evolved to replicate typological patterns seen in modern languages. Results in [1] show the emergence of universal colour categories in independent populations based on simulations using data from the WCS. In addition, Belpaeme and Bleys [3] investigated the roles of language interaction and environmental colour distribution in language games. They found that the best convergence on the WCS data was achieved with a uniform distribution on colour data and where agents were allowed language interactions.

In this paper we present a variation on the guessing game where categories are defined in terms of prototypes and also embed a measure of semantic uncertainty. Semantic uncertainty [31] is the uncertainty associated with the definition of concept labels, and results from the empirical manner in which we all learn to use language. As such it can capture natural differences between individuals about how a label is interpreted, as well
as the inherent subjective uncertainty that each individual language user has about the actual definition of the label. Care should be taken to distinguish semantic uncertainty from *stochastic uncertainty* which refers to uncertainty about the actual (or future) state of the world. This uncertainty not only allows different agents to assign different meanings to a concept, but also allows an individual agent to define categories in an imprecise manner, in which there is inherent uncertainty as to where each concept boundary lies. In keeping with the language game literature, agents are ignorant about other agents’ cognitive models. However, concepts are not modelled as partitions of the unit interval as in [1] but rather as uncertain convex regions of a metric space. The number of concept labels used by each agent is fixed and the evolution of the system is measured in terms of the change to these underlying convex regions during the game. Furthermore, we propose a richer model of concepts, and a richer model of language assertions which include reference to negated labels and, in principle, other compound logic expressions. Updating algorithms are defined based on minimal cognitive changes to satisfy the constraints introduced by such assertions. This richer representational structure allows for a more flexible approach to categorisation, incorporating aspects of typicality and semantic uncertainty. Furthermore, in contrast to Gärdenfors [19] and Douven et al [13] we do not insist that concept labels form an exact partition of the conceptual space. Instead, we expect that labels will overlap to some extent and also that there may be regions of the space to which no label is applicable.

3 Modelling Concepts

3.1 Conceptual Spaces and Prototype Theory

Categorisation is a fundamental part of human cognition. Usually when we encounter objects, some form of categorisation process is performed, requiring humans to constantly make decisions about how best to describe the world around them. The classical theory of categorisation requires that members of a given category all satisfy a set of shared properties. However, psychologists have criticised this model arguing that a richer, more flexible representation is required for natural categories (see [29] for an exposition). The classical requirement that category definition be characterised in terms of a finite set of necessary and sufficient conditions makes categorisation a highly inflexible process in practice. For instance, this approach tends to prevent the inclusion in a category of cases which technically violate the category definition but are none-the-less very similar to elements belonging to the category. For example, birds are typically described as ‘animals which fly’, but there are members of the category ‘bird’ which cannot fly, such as the ostrich or the penguin.

Prototype theory is an alternative model of category representation which has been proposed by Rosch [44, 43, 42] and developed by Lakoff [29] and more recently by Hampton [23]. Prototype theory proposes that categories are defined by similarity to a set of prototypical cases, instead of a set of necessary and sufficient conditions. This model naturally results in a typicality ordering on possible examples of a category. Furthermore, [44] presents evidence that typicality effects are not restricted by culture or semantic coverage. In other words, categorisation transcends cultural differences as a process of comparison to
a best-representative(s) of a category, i.e. a prototype or set of prototypes. In this context, a prototypical representation of a concept need not necessarily be fully grounded in the structure of a concept representation but may be more of an abstract notion [23].

A measure of conceptual distance (or similarity) is a fundamental aspect of prototype theory. In many natural categories there is an inherent measure according to which some category members are more similar than others. For example, a sparrow is more similar to a starling than to an ostrich, although all three are members of the category “bird”. Similarity spaces have been an active topic of discussion over the last several decades (e.g. [39, 46]), and more recently Gärdenfors has proposed conceptual spaces [19, 18] as a geometrical framework for concept representation. A conceptual space is a (multi-dimensional) metric space where each dimension quantifies a certain property or feature, so that a given object is fully described by a point in this space. The conceptual spaces framework theoretically permits a description of objects from any given category by selecting measurable features for the quality dimensions of a conceptual space. In practice not all concepts are geometrically measurable, however some geometrical spaces exist which adhere to the conceptual spaces framework. For example, CIELAB is a perceptually accurate geometrical colour space. This perceptual accuracy gives a relation between a change in CIELAB space (in terms of geometric distance between points in the space) and a perceptual difference perceived in these colour representations.

Categories within a conceptual space are represented as convex regions, and in the case that these form a partition of the space, then the set of labels could be generated by a Voronoi tessellation based around prototypical points [42], and hence providing a natural link to prototype theory. Indeed, in the context of language games Jäger, and Jäger and Van Rooij, [27, 28] have shown that a signaling game manipulating agents’ definitions of colours in RGB space can converge to a model equivalent to a Voronoi tessellation of prototypical colour points under general assumptions. Defining such a tessellation on a conceptual space, however, will restrict an agent to having a set of categories which are mutually exclusive and exhaustive, as assumed a priori by Fine [17]. From this perspective Voronoi conceptual models provide no obvious explanation or motivation for the use of no atomic statements such as negations or conjunctions. Why describe a colour as not red when there is always an appropriate colour label to describe it. Here we propose an alternative neighbourhood based definition of categories where each category boundary may be determined independently from other category prototypes. This allows for concepts to overlap within the space and also for there to be regions of the space in which no category is defined (see figure 1). This latter property in particular provides a clear motivation for the use of negated statements in certain contexts.

3.2 A Random Set and Prototype Theory Approach

Vagueness is a ubiquitous phenomenon in natural language [45], inherent in most lexical categories. As distinct from ambiguity 1 vagueness refers to the situation in which a lexical structure does not have a unique interpretation. For example, consider the word “tall”. Exactly what meaning is conveyed by declaring a person to be tall? Over 5’8, over 6 feet, between 6 feet and 6’3, exactly where does the boundary between tall and not tall

1i.e. where there are multiple linguistic meanings
lie?

In [12], van Deventer proposes a number of ways in which vagueness may be a positive feature of language. For instance, he suggests that risk may be minimised by using vague assertions in the presence of conflict using the example of two politicians campaigning for election. He argues that making firm commitments can be a poor strategy for politicians, risking over-commitment to the electorate which may in turn result in diminished popularity or standing when these commitments are not fulfilled. A vague commitment, however, is a lower risk strategy for a politician, leaving no exact thresholds to satisfy whilst still giving a promise to the electorate. For example, “our policies will result in a significant reduction in youth unemployment” is a lower risk assertion than promising “we will reduce unemployment for people between 18 and 24 by at least 75%.”

Given the distributed manner in which language is learnt, agents are unsure as to the ‘correct’ definition of categories and also how different individuals will interpret labels. By explicitly representing that uncertainty we can make better judgments and decisions about how best to describe the world. The epistemic theory of vagueness presented by Williamson [58] assumes the existence of an objectively correct set of criteria for determining whether or not a given instance satisfies a vague concept. Furthermore, Williamson argues that such criteria cannot be completely known to individuals resulting in semantic uncertainty. We would further suggest that, in fact, the conventions of language are not determined by some external authority but are best represented by a distributed body of knowledge across a population. Previous results within the language games literature would support this hypothesis. On the other hand, in practice, individual agents may find it useful as part of a decision making process to assume that such a set of rules does actually exist. That is to say, each agent will behave as if the epistemic theory of vagueness is correct. This pragmatic assumption is referred to as the epistemic stance [33].

In this paper we use the framework introduced by Lawry et al. [32, 33] to model concepts based on a combination of prototype theory and random set theory. This framework allows us to embed concepts within a conceptual space, and to capture elements both of typicality and semantic uncertainty. Note that prototype theory has already been applied in guessing games. For example, a prototype approach was adopted in [3] within their category model, where the membership function of a category was taken to be the inverse of
the Euclidean distance to the prototype. This model was criticised in [41] on the grounds that you can define an equivalence relation between this and their rigid boundary model. However, this is not the case for our neighbourhood based model as we are not using prototypes simply as a means of generating a partition of the conceptual space according to a Voronoi tessellation.

A Voronoi-based cognitive model has been proposed by Douven et al [13] accounting for category vagueness by considering the underlap and overlap of each category boundary. The explicit representation of borderline cases is then achieved by modelling each prototypical point as an uncertain region, naturally generating uncertainty over category boundaries. This approach, however, still means that the categories form an exhaustive covering of the conceptual space. Examples of implementations of language games using prototype or exemplar-based methods which do not use a Voronoi tessellation are given by Baumeister [2] and Spranger and Pauw [48]. Baumeister [2] assumes a priori that agents have dynamic categories, in that they will adapt category boundaries based on interaction with other agents i.e. as in language games. Following this assumption the author presents a proof that under certain conditions agents cannot determine a sharp category boundary. Spranger and Pauw [48] investigate language games in robotics, involving vague categories. Here the problem of grounding language in sensorimotor spaces is considered, and a degree-based approach to vagueness is proposed as a model of categories. In section 3.3, we present a conceptual model which allows us to define concepts by uncertain regions of a conceptual space. Under this definition concepts have boundaries which are represented by a random variable (an uncertain distance). Furthermore, categories may overlap with other categories and there may be regions of the conceptual space where no category is defined. In other words, our approach leaves the categorisation incomplete and if we consider only assertions of basic labels such as “the colour of desk is brown”, then there would tend to be regions of the conceptual space in which agents would not be able to identify any appropriate assertion. However, in contrast to Baumeister [2] and Spranger and Pauw [48] we allow negated assertions which can be used to describe regions of the conceptual space with low membership. In practice, there is the potential to locate prototypes and select thresholds depending on the distribution of elements in the conceptual space as governed by the environment. In this way, the effect of the incomplete coverage of the space provided by the neighbourhood model can be reduced if labels neighbourhoods are defined so that those elements lying outside the scope of all the labels have relatively low probability of actually occurring.

More generally, the neighbourhood prototype model also opens the possibility of conjunctions and other compound descriptions being used as a means to provide more precise information as and when appropriate. In particular, conjunctions of labels can be useful to describe elements which lie on the borderline between one label and another. For example, consider a search problem in which an individual is searching for a particular book on her friend’s bookcase. She phones her friend who describes the colour of the book’s cover to her. If this colour lies on the borderline between yellow and orange then the description yellow-orange, or yellow ∧ orange, is likely to be much more useful during the search than either of the single label descriptions of yellow or of orange. Yet conjunctions of labels can play no role in Voronoi-based models since by definition the labels refer to distinct

\footnote{Note there is evidence that using negative assertions can increase argument strength [25] and also that both positive and negative assertions are required to delineate a concept [57].}
and non-overlapping regions of the conceptual space. Indeed, there is no way of explicitly referring to borderline elements which whilst lying within the scope of one label are close to the borderline with another.

3.3 Representing Concept Labels in Conceptual Space

Suppose we have an underlying conceptual space which we denote by $\Omega$ and an associated distance metric $d$. In the following experiments we will take $\Omega$ to be the unit cube $[0,1]^3$ and $d(x,y) = \|x - y\|$ to be the standard Euclidean metric. We now define a finite set of labels $LA = \{L_1, \ldots, L_n\}$ where each label $L_i$ represents a word which can be used to describe elements of the conceptual space $\Omega$. In this case basic labels $L_i$ correspond to descriptive words (i.e. adjectives or nouns) for which the expressions ‘$x$ is $L_i$’ are meaningful for any $x \in \Omega$. For example, taking $\Omega$ to be a colour space such as rgb or CIELAB then $LA$ could consist of the basic colour labels red, orange, blue etc. We also consider negative statements in the form of negated labels e.g. not red for which we use the standard logic notation $\neg L_i$, so that the description of element $x$ by negated label $\neg L_i$ corresponds to the assertion that ‘$x$ is not $L_i$’ or alternatively that ‘$x$ cannot be described as $L_i$’.

As outlined above we will interpret the meaning of labels by using the neighbourhood prototype model. In this way each label $L_i$ is defined by a prototypical element of $\Omega$ denoted $p_i$, and a distance threshold $\epsilon_i \geq 0$. The intuitive idea is that an element $x$ of $\Omega$ can be described using label $L_i$ if and only if it is close to the prototype $p_i$. Closeness is then modelled by the requirement that the distance between $x$ and $p_i$ is less than or equal to the threshold $\epsilon_i$ i.e. $d(x,p_i) \leq \epsilon_i$. In order to capture the vagueness associated with natural language terms we assume that the thresholds $\epsilon_i$ are random variables each with an associated density function denoted by $f_i$ (see figure 2). From this perspective the meaning of $L_i$ is represented by an uncertain neighbourhood of the prototype $p_i$ and hence forms a random set of the space $\Omega$ [38] [36]. This approach naturally allows us to define membership functions for each label where $\mu_{L_i}(x)$ quantifies the extent to which element
$x$ satisfies or can be described by the label $L_i$. In contrast to fuzzy set theory [59], in the
eighbourhood model membership has an entirely probabilistic interpretation according to
which $\mu_{L_i}(x)$ is the probability that $x$ is close to prototype $p_i$. More formally, if $F_i$ is the
cumulative distribution function of $f_i$, and $CF_i = 1 - F_i$ is the complementary cumulative
distribution function then:

$$\mu_{L_i}(x) = P(d(x, p_i) \leq \epsilon_i) = 1 - F_i(d(x, p_i)) = CF_i(d(x, p_i))$$

Similarly, according to the neighbourhood model the negated label $\neg L_i$ can be used to
describe element $x$, provided that $x$ is not close to $p_i$ i.e. $d(x, p_i) > \epsilon_i$. Consequently the
membership of $\neg L_i$ is given by:

$$\mu_{\neg L_i}(x) = P(d(x, p_i) > \epsilon_i) = F_i(d(x, p_i)) = 1 - CF_i(d(x, p_i)) = 1 - \mu_{L_i}(x)$$

**Example 1.** Let $\Omega = \mathbb{R}^k$ and $d$ be the standard Euclidean distance given by $d(x, y) = ||x - y||$. Suppose that $f_i$ is the uniform distribution on the interval $[0, b_i]$ for some $b_i > 0$
so that

$$f_i(\epsilon) = \begin{cases} \frac{1}{b_i} : 0 \leq \epsilon \leq b_i \\ 0 : \text{otherwise} \end{cases} \quad \text{and} \quad CF_i(d) = \begin{cases} \frac{b_i - d}{b_i} : 0 \leq x \leq b_i \\ 0 : \text{otherwise} \end{cases}$$

Hence, the membership functions of $L_i$ and $\neg L_i$ are given by:

$$\mu_{L_i}(x) = CF_i(||x - p_i||) = \begin{cases} \frac{b_i - ||x - p_i||}{b_i} : ||x - p_i|| \leq b_i \\ 0 : \text{otherwise} \end{cases} \quad (1)$$

and

$$\mu_{\neg L_i}(x) = 1 - \mu_{L_i}(x) = \begin{cases} \frac{||x - p_i||}{b_i} : ||x - p_i|| \leq b_i \\ 1 : \text{otherwise} \end{cases} \quad (2)$$

Figure 3 illustrates the simple case in which the conceptual space is the real numbers, the
prototype $p_i$ is 10 and $f_i$ is the uniform distribution between 0 and 5 i.e. where $b_i = 5$.

So for the neighbourhood model we can think of an agent’s interpretation of the labels
$LA$ as being characterised in a certain metric space $\Omega$ with distance $d$, by the prototypes
$p_i$ and threshold random variables $\epsilon_i$ together with their associated density functions $f_i$
for each of the different labels $L_i$. These interpretations are not static but instead are
regularly updated as a result of communication between agents. In the sequel

**Definition 2. Interpretation**

An interpretation of the label set $LA$ is a tuple $I = (\Omega, d, \bar{p}, \bar{c})$, where $\Omega$ is a conceptual
space and $d$ is the associated metric, $\bar{p} = (p_1, \ldots, p_n)$ is the vector of prototypes, and
$\bar{c} = (\epsilon_1, \ldots, \epsilon_n)$ is the vector of threshold random variables where $\epsilon_i$ has the density function
$f_i$ for $i = 1, \ldots, n$.

In the sequel, when $I$ is dynamic and in particular when we describe rules for updating
an interpretation, we will use the notation $\mu_{I_i}^I(x)$ to denote the membership function of
$L_i$ in interpretation $I$. In the case that $I$ is fixed we will drop the superscript and simply
write $\mu_{L_i}(x)$. 

Figure 3: For the space $\Omega = \mathbb{R}$ and label $L_i$ with prototype $p_i = 10$ and $f_i$ corresponding to the uniform distribution on the interval $[0, 5]$. Figure 3 (a) shows the density function $f_i$. Figure 3 (b) shows the complementary cumulative distribution function $CF_i$. Figure 3 (c) shows the resulting membership function $\mu_{L_i}(x)$.

4 Concept Evolution in a Language Game

4.1 The Language Game Model

We consider a population of $N$ agents, each of whom has defined a set of basic labels in a conceptual space $\Omega$ modelled using the prototype neighbourhood representation as described in section 3.3. We take $\Omega = [0, 1]^3$, and $d$ to be the Euclidean norm. Each agent has an associated weight $w \in [0, 1]$ quantifying the level of importance other agents give to their assertions. The simulations progress in discrete time steps, and at each time step every agent interacts with another in a pairwise manner as part of a language game, where one agent plays the role of a speaker and the other a listener. This results in $N/2$ interactions at each time step. As in example 1 we will assume that for each label $L_i$ the threshold $\epsilon_i$ is distributed according to a uniform distribution on the interval $[0, b_i]$ for some positive real value $b_i$.

For each interaction both agents are presented with an element $x$ from the conceptual space $\Omega$. The speaker then asserts ‘$x$ is $\theta$’ to the listener where $\theta$ is either $L_i$ or $\neg L_i$ for some label $L_i$ in $LA$, as chosen according to the assertion algorithm described in the next section and based on their current interpretation $I$. The listener then calculates the membership $\mu_\theta(x)$ according to their interpretation of $LA$ at the given time step. In the case that $\mu_\theta(x) \geq w$, where $w$ is the speaker’s weighting, then the listener takes no action, since they already believe the relevant assertion at least to the level of the weight of the speaking agent. On the other hand, if the membership $\mu_\theta(x) < w$ then they update their interpretation of the labels so that the inequality $\mu_\theta(x) \geq w$ does hold. We assume that for

---

This choice of metric space was originally inspired by other work done on colour spaces such as the RGB cube. Here, while we are not focusing on a specific application, considering a higher-dimensional space allows for a more general investigation of the proposed framework than, for example, the real line [1].

The weight $w$ can be thought of as the age or seniority of an agent. The reasoning behind this is that in a conversation one might be inclined to put more faith in someone who is older and has more experience.
Figure 4: A graphical representation of a pairwise language game interaction in which $w$ is the weight associated with the speaking agent.

The guessing game agents form a fully connected network in which all pairwise interactions are equally possible. Then at each time step we uniformly randomise the set of agents so that half are speaking agents and half are listening agents\(^5\) (see figure 4).

4.2 The Assertion Model

For a given element $x$ in the space $\Omega$, an agent has available as possible assertions any label $L_i$ which is close to $p_i$ and any negated label $\neg L_i$ which is not close to $p_i$. However, given the uncertainty associated with the distance thresholds $\epsilon_i$ there is naturally uncertainty as to what constitutes this set of assertable descriptions of $x$. Let the set of assertable labels and negated labels for $x$ be denoted by:

$$\text{AS}_x = \{L_i : d(x, p_i) \leq \epsilon_i\} \cup \{-L_i : d(x, p_i) > \epsilon_i\}$$

So, for example, if $LA = \{L_1, L_2, L_3, L_4\}$ then the probability that the set of possible assertion is $\text{AS}_x = \{L_1, L_2, \neg L_3, \neg L_4\}$ is the joint probability that $\epsilon_1 \geq d(x, p_1)$, $\epsilon_2 \geq d(x, p_2)$, $\epsilon_3 < d(x, p_3)$ and $\epsilon_4 < d(x, p_4)$ (see figure 5). To determine this probability requires knowledge of the relationship between the different threshold random variables, and in the worst case a full joint density on $\epsilon_1, \ldots, \epsilon_4$. In this paper we make an assumption of strong dependence between the thresholds which allows us to determine the required probabilities directly from the membership values for $x$. In particular, we assume that there is a single underlying threshold variable $\epsilon$ which all of the thresholds are proportional to i.e. that $\epsilon_i = k_i \epsilon$ for some positive constant $k_i$. Based on this assumption we can use the following simple algorithm to determine the probabilities for the different possible assertion sets $\text{AS}_x$ [33]: Let the labels be ordered according to increasing membership so that $\mu_{L_1}(x) \geq \mu_{L_2}(x) \ldots \geq \mu_{L_n}(x)$. Then the only possibilities are that $\text{AS}_x = \{L_1, \ldots, L_i, \neg L_{i+1}, \ldots, \neg L_n\}$ for some $i = 1, \ldots, n$. Furthermore, in each case the probability is:

$$P(\text{AS}_x = \{L_1, \ldots, L_i, \neg L_{i+1}, \ldots, \neg L_n\}) = \mu_{L_i}(x) - \mu_{L_{i+1}}(x)$$

\(^5\)Note that there is evidence to suggest that language networks are appropriately modelled by small world networks [9].
So for our previous example, assuming that \( \mu_{L_1}(x) \geq \mu_{L_2}(x) \geq \mu_{L_3}(x) \geq \mu_{L_4}(x) \) then

\[
P(\text{AS}_x = \{L_1, L_2, -L_3, -L_4\}) = \mu_{L_2}(x) - \mu_{L_3}(x)
\]

We now assume that for every element \( x \), each agent believes that there is a most appropriate assertion to make describing \( x \). Let this be denoted by \( A_x \). The above method can now be combined with standard Bayesian reasoning in order to determine a probability distribution on \( A_x \) taking account of the agent’s interpretation \( I \) of the labels. Initially we assume a prior distribution on \( A_x \) given by:

\[
P(A_x = L_i) = \frac{q}{n} \quad \text{and} \quad P(A_x = -L_i) = \frac{1-q}{n} \quad \text{for } i = 1, \ldots, n
\]

Here the parameter \( q \) quantifies the agent’s prior propensity to choose either positive or negative assertions. Otherwise all labels are a priori equally probable, as are all negated labels. Based on this prior we can now define a posterior distribution on assertions, conditional on the agent’s interpretation of the labels. If the set of all possible assertions to describe \( x \), \( \text{AS}_x \), was know with certainty then this posterior distribution would simply be given by:

\[
P(A_x | \text{AS}_x) = \begin{cases} 
P(A_x) & : A_x \in \text{AS}_x \\ 0 & : \text{otherwise} \\ \end{cases}
\]

where \( P(A_x) \) and \( P(\text{AS}_x) \) are determined from the prior distribution. However, as discussed above \( \text{AS}_x \) is uncertain and we take this into account by applying the law of total probability so that:

\[
P_I(A_x) = \sum_{\text{AS}_x} P(A_x | \text{AS}_x)P(\text{AS}_x)
\]

In practice we make a minor adaptation to this assertion model in that we only allow the assertion of negated labels where the label itself has non-zero membership value. Whilst, formally, negations of labels with zero membership constitute valid assertions, preliminary investigations suggested that making such assertions tends to skew the updating process and reduce the level of convergence towards a shared interpretation between agents. Furthermore, these assertions would seem to lack relevance in the context of a language game in which the speaker is attempting to describe \( x \) so as to convey information about her interpretation of \( LA \) to the listener. For example, consider a colour space described by a standard set of colour labels. Suppose that the speaker has some non-zero belief less than one that a colour \( x \) can be described as \( \text{red} \), but she is certain that all other labels are inappropriate i.e. \( 0 < \mu_{\text{red}}(x) < 1 \) and \( \mu_{L_i}(x) = 0 \) for \( L_i \neq \text{red} \). In this case, despite the fact that \( 1 = \mu_{\text{black}}(x) > \mu_{\text{red}}(x) \) we argue that in a language game the assertion ‘\( x \) is not red’ is more informative than the assertion ‘\( x \) is not black’. Underlying this intuition is the fact that \( x \) is a \textit{borderline case} of \( \text{red} \), since both \( \mu_{\text{red}}(x) > 0 \) and \( \mu_{\text{not-red}}(x) > 0 \), but it is not a borderline case of any other colour. The implicit assumption then is that it is particularly with regard to borderline cases that agents tend to differ in their interpretation of labels, and hence focusing assertions on such cases will tend to result in better convergence towards a common shared interpretation across the population.
Figure 5: An example of prototypes and thresholds for $L_1, \ldots, L_4$ where the set of possible assertion to describe $x$ is $\text{AS}_x = \{L_1, L_2, \neg L_3, \neg L_4\}$.

4.3 The Updating Algorithm

We now focus on the second part of an interaction between agents in which the listener updates their interpretation $I$ based on the assertion made by the speaker. Consider the assertion ‘$x$ is $\theta$’, where $\theta$ is $L_i$ or $\neg L_i$, made by a speaker with weight $w$. In our language game the listener is then required to update their interpretation of $LA$ from $I$ to $I'$, in such a way that the membership constraint $\mu^I_\theta(x) \geq w$ is now satisfied. In the case that for the current interpretation $I$, $\mu^I_\theta(x) \geq w$ already holds then no update is required and the listener takes $I' = I$. Since we are permitting both positive and negative assertions by the speaker then we must consider the following two cases (see figure 6):

1. Consider the positive assertion ‘$x$ is $L_i$’ in an interaction where the listener’s interpretation $I$ is such that $\mu^I_i(x) < w$. In this case we propose that the listener updates their interpretation from $I$ to $I'$ by changing their interpretation of the label $L_i$ only. In particular, we propose to update the prototype of $L_i$ from $p_i$ to $p'_i = p_i + \lambda(x - p_i)$ where $\lambda \in [0, 1]$, and the threshold for $L_i$ to $\epsilon'_i = \alpha \epsilon_i$ for $\alpha \geq 1$. The update to $p_i$ moves the prototype closer to $x$, hence reducing the distance between $x$ and the prototype so that the membership value increases. In terms of the update to the threshold, recall that we are assuming a uniform threshold distribution on the interval $[0, b_i]$. Consequently, the scaled threshold random variable has a uniform distribution on the interval $[0, \alpha b_i]$ and since $\alpha \geq 1$ this also results in an increase in the membership values for all $x$ in the space.

2. Consider the negative assertion ‘$x$ is $\neg L_i$’ in an interaction where the listener’s interpretation $I$ is such that $\mu^I_\neg L_i(x) < w$ (i.e. $\mu^I_i(x) > 1 - w$). Again we propose that the listener updates their interpretation from $I$ to $I'$ by changing their interpretation of the label $L_i$ only. In particular, we propose to update the prototype of $L_i$ from $p_i$ to $p'_i = p_i - \lambda(x - p_i)$ for $\lambda \in [0, 1]$ and the threshold for $L_i$ from $\epsilon_i$ to $\epsilon'_i = \alpha \epsilon_i$
for $0 < \alpha < 1$. The update to $p_i$ moves the prototype away from $x$, hence increasing the distance between $x$ and the prototype so that the membership value decreases. Consequently the membership of $\neg L_i$ increases. In addition, the scaled threshold random variable now has a uniform distribution on the interval $[0, \alpha b_i]$ which results in a decrease in membership for all $x$ in the space, since $0 < \alpha < 1$.

There remains the question of how to determine the parameter values $\lambda$ and $\alpha$ for the updating rules described above. Clearly these should be chosen so that the constraint imposed on the asserted label is met i.e. $\mu_{L_i}(x) \geq w$ for a positive assertion and $\mu_{L_i}(x) \leq 1 - w$ for a negative assertion. However, this requirement is not in general sufficient to identify unique values for $\lambda$ and $\alpha$. Therefore, we introduce an additional minimal updating assumption according to which the listener makes the minimal change to their interpretation of the labels so as to satisfy the constraints on membership values imposed by the speaker’s assertion. To quantify the difference between the speaker’s original and updated interpretations we use the Hausdorff metric. This is a standard metric for measuring the distance between two sets within a metric space. In the current context we measure the Hausdorff distance between the listener’s original and updated neighbourhoods for the label $L_i$. For any fixed threshold value $\epsilon_i$ there are unique values for $\lambda$ and $\alpha$ so that the updated interpretation satisfies the required membership conditions whilst minimizing the distance between the updated and original neighbourhoods (see Eyre [16] for details). Taking the expected value as $\epsilon_i$ varies then results in the following parameter settings:

Given a positive assertion

$$\lambda = w(1 - \frac{b_i(1 - w)}{\|x - p_i\|}) \quad \text{and} \quad \alpha = \frac{\|x - p_i\|}{b_i} + w$$
Given a negative assertion

\[ \lambda = (1 - w)\frac{b_i w - \|x - p_i\|}{\|x - p_i\|} \]  
and \[ \alpha = \frac{\|x - p_i\|}{b_i} + 1 - w \]

5 Measures of Convergence and Overlap

In order to evaluate the performance of the above assertion and updating algorithms we now consider the degree to which agents agree across the population. A natural quantification of this level of agreement is given by the distance between agent interpretations. In addition, we also propose a measure of the degree to which an individual agent’s categories overlap. In the above we have argued that a conceptual model in which categories overlap is more expressive. However, this should be balanced against the agent’s need to use category labels as a way of discriminating different regions of the conceptual space. For example, we conjecture that too much overlap between colour labels would reduce the effectiveness of colour as a way of discriminating between different coloured objects.

We now introduce two performance metrics, average pairwise distance between interpretations (APDI) and average pairwise overlap between labels (ALO), to quantify convergence and overlap respectively. APDI uses the Hausdorff metric to measure the distance between the interpretations of two agents. This is calculated for every pair of agents in the population and then averaged across all pairs. In addition, for a given element \( x \) the degree to which \( x \) belongs to both labels \( L_i \) and \( L_j \) can be determined by taking the minimum of the membership values \( \min(\mu_{L_i}(x), \mu_{L_j}(x)) \). The maximum of this expression across the elements of the space then provides a measure of the extent to which labels \( L_i \) and \( L_j \) overlap in a given interpretation.

\[ \max\{\min(\mu_{L_i}(x), \mu_{L_j}(x)) : x \in \Omega\} \]

Averaging across all pairs of labels quantifies the level of overlap between different labels in a given interpretation. This is then averaged again over the interpretations of all the agents in the population to give ALO.

6 Experiments

Experiments were carried out using the algorithms described in section 4. Each simulation of was run using a population of 100 agents for 10000 time steps on a fully connected network. At each time step every agent was assigned an interacting agent at random so that half the population were interacting as speakers, the other half as listeners, so at each time step 50 language games. Weights were initially assigned to agents on the interval \([0.1, 0.9] \) ensuring a uniform distribution of weights across the population.

Agents’ interpretations were defined by 11 prototypes which were selected at random for each agent at time step 0 according to the uniform distribution on \([0, 1]^3 \). The label thresholds were characterised by the vector \((b_1, \ldots, b_n) \). This defined the distribution of

---

\(^6\)Early experiments indicated non-stable behaviour at \( w = 0 \) and \( w = 1 \).
boundary variable \(\epsilon_i\) as the uniform distribution on \([0, b_i]\). Each \(b_i\) was initialised at random from the interval \([0.5, 2]\).

At each time step the weight of each agent was increased by \(8 \times 10^{-6}\), until it reached 0.9 at which point it was reset to 0.1. This models a simplified ageing process, in which agents are ‘re-born’ at their maximum age, so that a new-born agent inherits the interpretation \(I\) from the dying agent, perhaps as human children may expect to inherit a set of concept definitions initially similar to those of their parents.

In sections 6.1 and 6.2 we present results from a set of simulations testing different assertion priors i.e. \(q\) values. In all experiments the conceptual space \(\Omega\) is taken to be \([0, 1]^3\), and elements \(x\) to be described in each language game are chosen uniformly from across \(\Omega\). All results presented are averaged over 25 runs.

### 6.1 Results of APDI and ALO for different assertion priors

Figure 7(a) shows a decrease in APDI for simulations using both priors of \(q = 0\) and \(q = 1\). This indicates that we have a convergence in agent interpretations in both systems. Furthermore, we see a much greater convergence in APDI when we take \(q = 1\), i.e. when assertions of the type ‘\(x\) is \(L_i\)’ are almost exclusively preferable. Figure 7(b) shows that when we use a prior of \(q = 0\), ALO decreases, and increases when \(q = 1\). This is not a surprising result as the updating rule for positive assertions increases the boundary distribution parameter \(b_i\).

Regarding the results for \(q \in [0, 1]\) very similar behaviour was observed for the range \(q \in [0, 0.5]\), and also for \(q \in [0.6, 1]\), with a transition occurring between these intervals. To test for observable transitions in the behaviour of the system we took our averaged results and plotted the value observed for APDI and ALO at the end of each simulation against the assertion prior \(q\) used in the simulations. The results for the range \(q \in [0.5, 0.6]\) are shown in figure 8.
6.2 Communication Games

In the previous section we have postulated that APDI and ALO are good indicators of the communicative efficacy of the category labels LA. However, we have yet to investigate this claim directly. In this context communicative success requires that agents agree on the choice of assertions describing the elements of $\Omega$. Here we propose a form of communication game, in order to quantify how consistently a set of categories can be used between agents, as follows:

At each time step, once the pairwise guessing game has taken place, we assign each agent a different partner at random from the population. Each pair of agents are presented with 10 elements selected uniformly from $\Omega$. Both agents then select the label with maximal appropriateness measure according to their respective category models, for each of the 10 elements. The proportion of elements for which the agents agree on the most appropriate label, then provides a measure of communication success for this pair of agents. We then take the average of this measure over all agent pairs to give a measure of population communication success.

Investigations of simulation performance in terms of population communication success against time showed an increase in value for all assertion priors tested. For $q = 0$, where negated assertions are primarily used, we see a very small increase in population communication success as shown in figure 9(a)$^7$. Simulations using values of $q > 0.5$ result in much larger increases in population communication success. As with APDI and ALO, by considering the value for population communication success at the end of simulations, it is easier to see which values of $q$ converge to the highest values of population communication success. Figure 9(b) shows these results for assertion priors in the range $q \in [0.5, 0.6]$. The assertion prior $q = 0.545$ clearly gives the highest value for population communication success at time step 10000. If we consider the plot of population communication success

---

$^7$At time step 0 we observe a population communication success of about 0.1. This is roughly equivalent to picking a label at random from a set of 11 defined for agents’ interpretations.
Figure 9: Figure 9(a) shows the population communication success against time for simulations where $q = 0$, $q = 0.545$, and $q = 1$. Figure 9(b) shows the value for population communication success at time step 10000 for simulations for priors where $q \in [0.5, 0.6]$.

against time shown in figure 9(a) we see that the result for this optimal assertion prior ($q = 0.545$) and the prior of $q = 1$ are clearly distinguishable. This suggests that a mixed assertion model incorporating both positive and negative assertions gives the best performance in terms of population communications success. From figure 8 we see that setting $q = 0.545$ results in good category convergence across the population (i.e. a low APDI value), whilst maintaining some discrimination between categories (i.e. a non-maximal ALO value). Such a trade-off may provide some explanation for the relative efficacy of a mixed assertion model, over a model which only allows for the positive assertion of category labels.

7 Conclusions

We have presented a new framework for category language games, in which categories are represented in a conceptual space according to a random set and prototype theory model. More specifically, categories are defined by a prototype and a boundary determined by an uncertain distance threshold to a prototype. We have argued that this provides a flexible and expressive model of categories, and by explicitly representing semantic uncertainty agents are better equipped to make good decisions about how to describe the world. Using the neighbourhood conceptual model we have introduced algorithms for generating assertions, and for category updating. Updating algorithms involve a mixture of updates to both the prototypes and boundary threshold distributions. In this context we have identified updating rules which result in the minimal change to the listener agents’ interpretation which is required in order to satisfy a natural constraint imposed by the assertion and weight of the speaking agent. Here the difference between interpretations is quantified using the Hausdorff metric to measure the difference between neighbourhoods in the conceptual space.

In order to analyse the behaviour of the multi-agent system, we have introduced metrics
to measure overall convergence between interpretations (APDI) and the average label overlap between labels within an interpretation (ALO). Empirical results as obtained from applying these metrics in simulations of multi-agent systems where agents play the category language game have been presented. Results suggest that APDI converges to the lowest value when a mixed updating rule is applied, updating both the prototypes and the category boundaries. Considered together with the results for ALO, we may conclude that, for our particular experimental set-up, the optimal convergence in agent categories is obtained when the categories are not maximally overlapping.

This conjecture has then been investigated more directly through a type of communication game providing a method for measuring communication success independently of both the learning algorithms and the explicit comparison of agents’ conceptual models. Here we found that simulations using a mixed assertion set of both basic labels and their negations showed the greatest increase in population communication success.

As the prior probability of a positive assertion, q, tends to 1 we see a large increase in category overlap, which also tends to result in a more uniform assertion probability distribution across labels for all x in the space Ω. In effect, this will mean that agents will tend to believe that each category is equally appropriate to describe any given label. Whilst this ensures convergence of interpretation between agents, it does not allow for descriptions which can discriminate between different regions of the conceptual space. Instead, for communication success it is required that agents have converged on a similar set of category definitions (i.e. low APDI) which are (at least) not fully overlapping (i.e. ALO not maximal). Such a trade-off is fully consistent with an assertion model which mixes both positive and negative assertions.

Acknowledgments

Henrietta Eyre was supported by the EPSRC grant EP/E501214/1.

References


