INTERACTIVE CONCEPT ACQUISITION FOR EMBODIED ARTIFICIAL AGENTS

by

JOACHIM DE GREEFF

A thesis submitted to Plymouth University
in partial fulfilment for the degree of

DOCTOR OF PHILOSOPHY

School of Computing and Mathematics
Faculty of Science and Technology

July 2013

Interactive Concept Acquisition for Embodied Artificial Agents

by

Joachim de Greeff

Abstract

An important capacity that is still lacking in intelligent systems such as robots, is the ability to use concepts in a human-like manner. Indeed, the use of concepts has been recognised as being fundamental to a wide range of cognitive skills, including classification, reasoning and memory. Intricately intertwined with language, concepts are at the core of human cognition; but despite a large body or research, their functioning is as of yet not well understood. Nevertheless it remains clear that if intelligent systems are to achieve a level of cognition comparable to humans, they will have to posses the ability to deal with the fundamental role that concepts play in cognition.

A promising manner in which conceptual knowledge can be acquired by an intelligent system is through ongoing, incremental development. In this view, a system is situated in the world and gradually acquires skills and knowledge through interaction with its social and physical environment. Important in this regard is the notion that cognition is embodied. As such, both the physical body and the environment shape the manner in which cognition, including the learning and use of concepts, operates. Through active partaking in the interaction, an intelligent system might influence its learning experience as to be more effective.

This work presents experiments which illustrate how these notions of interaction and embodiment can influence the learning process of artificial systems. It shows how an artificial agent can benefit from interactive learning. Rather than passively absorbing knowledge, the system actively partakes in its learning experience, yielding improved learning. Next, the influence of embodiment on perception is further explored in a case study concerning colour perception, which results in an alternative explanation for the question of why human colour experience is very similar amongst individuals despite physiological differences. Finally experiments, in which an artificial agent is embodied in a novel robot that is tailored for human-robot interaction, illustrate how active strategies are also beneficial in an HRI setting in which the robot learns from a human teacher.

Contents

List of Figures

List of Tables

Acknowledgements

Obviously this thesis would not have been what it is without the help and influence of quite a few people. Varying from moral support, critical comments, scientific advise, vital distractions or just the odd chat, your input has been greatly appreciated.

First of all, Tony, you have been invaluable as a supervisor. As the initiator of the CONCEPT project you have been close to my scientific exploration which has culminated in this thesis. I very much appreciate all the help you gave me over the years; whether it was advise, scientific guidance, a technical discussion, some confidence boost I tended to need from time to time, or just a friendly chat, you have provided this.

Fred, as my partner in crime on the CONCEPT project, it has been quite a journey. I have appreciated working with you at home and abroad; your thoroughness, eye for detail, time management, appreciation of sound coding and humour have been mildly troubling on occasion, but also motivational, inspirational and above all hugely enjoyable. I'm sure you will get far, who knows where we will meet.

Many thanks to my examiners Dr Paul Vogt (external) and Dr Angelo Cangelosi (internal) for their thorough reading and examination of this thesis. Addressing the points they identified as needing correction has improved the thesis substantially.

Tony Morse and Paul Baxter, discussions with you have been inspirational; your scientific insights and theoretical knowledge have been a huge help. I have enjoyed working and writing with you.

I am also greatly indebted to my proofreaders. Frédéric Delaunay, Rachel Wood, Alex Smith, Paul Baxter and Bas de Haas, many thanks for your thorough examination of draft versions of my chapters; your input has greatly improved the text, as well as prevented some embarrassing spelling mistakes.

To the people from the ALIZ-E lab, Robin Read, Rachel Wood, Paul Baxter and Magdalena Leshtanska, I apologise for the countless times I interrupted your work flow and though processes by producing those hard-to-deny sounds of coffee preparation, along with the inevitable crunching sound of opening some kind of package. Your little lab has provided a haven of caffeine intake, sugar shots and intellectual stimulation throughout the years for which I am greatly indebted.

The inhabitants of PSQ B111 and the people from the Centre for Robotics and Neural Systems at Plymouth University, you have provided a stimulating and fruitful environment for scientific research; I've enjoyed the time I spend amongst you. Salomon Ramirez-Contla, Fabio Ruini, Martin Peniak, Zoran Macura, Chris Larcombe, Davide Marocco, Angelo Cangelosi, Francesca Stramandinoli, Marek Rucinski, Kristsana Seepanomwan, Alessandro Di Nuovo, Federico Da Rol, Frank Broz, Naveen Kuppuswamy, Giuseppe Filippone, Hande Ceilikkanat, Peter Gibbons, Chris Ford, Guido Bugmann, Anna-Lisa Vollmer and Nicholas Hemion, thank you for shaping this environment. Also the daily lunch sessions during which discussions would not shy away from any topic imaginable were legendary.

Thanks also to Elena Dell'Aquila, Joanne Clements, Lucy Cheetham and Carole Watson for providing help whenever I was lost in forms and regulations. You have made things quite a bit easier for me.

I am also grateful to get acquainted with the AlienHouse and its colourful inhabitants; it has providing shelter, a homely environment, great meals and of course countless BBQs and parties. Also thanks to all the members of the basketball team, you know who you are. It has been my pleasure to organise the game and play with you!

To my mother and father Gerrie & Ab, brother Alwin and sister Ragna; thank you for your support throughout these years. Close or far, you have always been there for me. To Bas, thank you for your friendship. We have been more or less following a similar path, with, as you rightly observed, some differences as well. To friends from the Netherlands, Mira, David, Manon, Robbert, Jeroen, Ewout, Peter and Daan, it was good to be able to visit you from time to time, even though it was infrequent. Hopefully we will get to spend more time in the future.

To Marieke. Thank you for your support and trust in me. Thank you for your love.

Author's declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award without prior agreement of the Graduate Committee.

This work has been carried out by Joachim de Greeff as part of the CONCEPT project (EPSRC EP/G008353/1) under the supervision of Dr. Tony Belpaeme.

Publications

de Greeff, J., Delaunay, F., and Belpaeme, T. (2012b). Active robot learning with human tutelage. In *Proceedings of the joint International Conference on Developmental Learning (ICDL) & Epigenetic Robotics 2012*, pages 1–6, San Diego, USA. IEEE

de Greeff, J., Baxter, P., Wood, R., and Belpaeme, T. (2012a). From Penguins to Parakeets: a developmental approach to modelling conceptual prototypes. In Szufnarowska, J., editor, *Proceedings of the Post-Graduate Conference on Robotics and Development of Cognition*, pages 8–11, Lausanne, Switzerland

de Greeff, J. and Belpaeme, T. (2011b). The development of shared meaning within different embodiments. In Triesch, J., editor, *Proceedings of the joint International Conference on Developmental Learning (ICDL) & Epigenetic Robotics 2011*, Frankfurt, Germany. IEEE

de Greeff, J. and Belpaeme, T. (2011a). Coordination of meaning within different embodiments through linguistic interactions. In *Alife Approaches to Artificial Language Evolution (AAALE), workshop at the 20th European Conference on Artificial Life (ECAL)*, Paris, France

Morse, A., de Greeff, J., Belpaeme, T., and Cangelosi, A. (2010). Epigenetic robotics architecture (ERA). *IEEE Transactions on Autonomous Mental Development*, 2(4):325–339

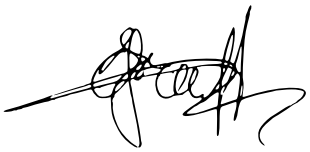
Delaunay, F., de Greeff, J., and Belpaeme, T. (2010). A study of a retro-projected robotic face and its effectiveness for gaze reading by humans. In *Proceeding of the 5th ACM/IEEE international conference on Human-robot interaction*, pages 39–44, Osaka, Japan. ACM/IEEE

de Greeff, J., Delaunay, F., and Belpaeme, T. (2009b). Human-robot interaction in concept acquisition: a computational model. In Triesch, J. and Zhang, Z., editors, *IEEE International Conference on Development and Learning (ICDL 2009)*, pages 1–6, Shanghai, China. IEEE

Delaunay, F., de Greeff, J., and Belpaeme, T. (2009). Towards retro-projected robot faces: an alternative to mechatronic and android faces. In *Proceedings of the International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Toyama, Japan

de Greeff, J., Delaunay, F., and Belpaeme, T. (2009a). Concept acquisition through linguistic human-robot interaction. In *Proceedings of the International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Toyama, Japan

Word count of main body of thesis: 44217

Signed: 

Date: July 2013

Chapter 1

Introduction

The use of concepts comes natural to people. We tend to talk and think in concepts, and the learning of new concepts or alterations of existing ones comes with remarkable ease. Concepts appear to be fundamental building blocks of human cognition. Hence, the ability to use concepts has been described as "paramount for the understanding of many cognitive phenomena" (Gärdenfors, 2000b, p.1), as all sorts of cognitive structures and abilities like memory, knowledge representation, reasoning and deduction entail concepts in some way. From ancient times scholars have been interested in this phenomenon, and particularly in the last century, along with the rise of modern psychology, the interest in concepts has increased tremendously. Yet the issue is far from settled, as despite a large body of scientific and philosophical investigations, there is still no consensus about what exactly constitutes a concept, how it might be encoded, how it can be acquired and in general what role it plays in human cognition.

Given the importance of this role in human cognition, it is not surprising that the topic has received considerable attention from artificial intelligence (AI) and cognitive robotics. This interest is not only in explaining the nature of human concepts, but also to enable artificial systems to use concepts in a human-like manner. The general aim of AI can thus be seen as twofold. On the one hand the development of AI models may help to understand human intelligence, as theories can be tested, made more concrete by requiring implementation details to be explicit and the application of models of cognition might create new predictions and provide new

inspirations. On the other hand, the fact that AI systems are currently far from intelligent might largely be due to a relatively poor understanding of human cognition. A more thorough understanding of human intelligence might turn out to be crucial if we ever hope to build artificial systems that display intelligent behaviour to some degree[1]. Thus, given the fact that the use of concepts is fundamental to human cognition, to build an effective AI system, a good understanding of how concepts function in humans seems necessary. And vice versa, to unravel the manner in which humans utilise conceptual knowledge, AI systems may prove to be invaluable tools. Both aims naturally go together, though for some researchers the first one may be the ultimate goal while for others it might be the latter, and some researchers might implicitly aim for both.

The interest in concepts from an AI perspective can also be found in Cognitive Robotics (CR), viewing the latter as an extension of AI that makes more rigorous demands with respect to methodology, design and robustness of effective AI systems. Indeed, due to the emphasis on real world applications and robotic embodiment, proposed solutions need to be able to cope with the, from an AI system's perspective, ever noisy[2], chaotic and dynamically changing environment that is constituted by the real world. At the same time this focus on embodiment, real world application and the increasing adoption of robots as part of society opens up new opportunities for interaction, learning and cooperation between robots and humans.

Setting the stage

The nature of the research described in this thesis is interdisciplinary, and as such themes from different disciplines, like psychology, AI and cognitive robotics will be brought forward. In the remainder of this chapter we will set the stage for the

[1]This, of course, depends on what exactly is meant with intelligent behaviour. The most famous 'definition' of artificial intelligence is formulated in the Turing Test (Turing, 1950), which states that an artificial system is said to be intelligent if its behaviour, while playing an imitation game with a human interrogator, cannot be distinguished from human behaviour. While this definition of intelligence is somewhat controversial (for instance, Searle's Chinese room thought experiment (Searle, 1980) argues against it), it is a fact that as of today AI systems have not reached a level of sophistication that enables them to pass something like a Turing Test, i.e. behave in such a manner that they cannot be distinguished from humans, other than in very restricted contexts.

[2]With 'noisy' the random fluctuations in sensory channels in general are meant, which are therefore not necessarily restricted to sound.

inclusion of these topics from different disciplines. We briefly discuss how different aspects from human concept use, AI, cognitive robotics, learning, developmental robotics, language and human-robot interaction all relate to the thesis at hand, that is, interactive concept acquisition for embodied artificial agents.

1.1 Concepts

One of the most compelling cognitive capacities is the ability to form concepts from observations we make in the world around us (Smith, 1995). A concept is generally understood to be some kind of abstract representation which somehow provides a 'model' for all the instances that we might encounter in the world around us; having such models allows for the sorting of streams of sensory information into meaningful 'chunks' as part of our cognitive processing. For now we will stick with this fairly informal description of concepts, a more in-depth discussion about the issues regarding a definition of concepts is provided in section 2.1 and section 2.2.

Having the capacity to use concepts provides a lot of advantages. For example, having a concept of a BEAR[3] allows us to recognise something with bear-like properties from a distance, make predictions about this (it might attack me if I get near) and communicate about this to others ("beware of the bear!"). There is no need to have encountered this particular bear ever before, and in fact it might look, smell or behave considerably different from other bears we have encountered, but yet we are able to do these things. Having the ability to match something we perceive with another thing that we have experienced before clearly serves a purpose. Yet, this is not a trivial task, as it requires a fine balance of sufficient but not too much detail in both the storage (memory) and matching process. If we have memorised too many details from previous encounters with bears, we might end up with a very specific BEAR concept and decide on that basis that the newly observed entity does not sufficiently matches to be considered a BEAR (a decision that could have lethal consequences). On the other hand, if too little is stored, our concept of BEAR would

[3]Throughout this thesis CONCEPTS will be written in small capitals to indicate the concept itself is meant (presumably as part of some agent's cognitive repertoire), rather than the linguistic label referring to this concept, or the actual referent in the world.

be very general and we might end up constantly warning people for all sorts of things that are brown and move. In a similar vein the matching needs to be balanced as well; our concept of BEAR might have struck the right balance between details and generality, but if we are, during the matching process, too fussy about the details of some observation or not fussy enough we still might be in trouble. Context is extremely important for this process (Barsalou, 1982), as this can determine which properties of a certain concept are relevant and which are not.

Humans are exceptionally good at using concepts in an effective manner. Not only visual observations from the world around us are subjected to conceptualisation, but virtually any sensory processing might be thought of in terms of concepts (Harnad, 1987). It would be very hard to formulate theories about cognition without resorting to a level of description that involve concepts in one way or another. However, despite large efforts to investigate and unravel human concept acquisition and use, no all-inclusive theory of concepts that manages to explain all empirical data has been proposed so far. While initially it seems that concepts, although central to human cognition, were relatively straightforward to study and to model, the more research was addressed to it, the more complex and 'messy' the story seemed to become (Murphy, 2002). Some authors have even proposed to stop trying to formulate theories of concepts altogether, arguing that their inclusion in cognitive theories only leads to misunderstandings (Machery, 2009). Nevertheless, despite the lack of a grand theory of concepts, it is an accepted fact that concepts are a fundamental part of human cognition, and are therefore of interest to anyone who wishes to study human intelligence[4].

1.2 AI and Cognitive Robotics

As previously described, the study of AI has two goals: understanding human cognition and the creation of artificial systems that exhibit human-like intelligence. Given

[4]Presumably concepts-like structures might be found in other forms of natural intelligence, see Allen and Hauser (1991) for a discussion. However, we are primarily concerned with human intelligence and concept use because the aim is to achieve human-like behaviour in artificial systems, rather than other animal-like behaviour.

the importance of concepts to human cognition, adopting either one of these goals entails an incorporation of mechanisms that can deal with human-like concept use. Traditionally however, AI has not necessarily focussed on modelling of concepts *per se*. Rather, it has undergone a kind of paradigm shift, going from a focus on modelling higher level cognitive functioning (like symbolic reasoning) to a much more basic level of embodied cognition (see Anderson (2003) for an overview of this shift). This transition is briefly described below.

1.2.1 From GOFAI to embodiment

The now famous Dartmouth conference in 1956 is generally credited as the birthplace of AI. While at the time cybernetics[5] already advocated a systems view, research in AI from the 1950s onwards tended to revolve around principles that are now generally dubbed as Good Old Fashioned Artificial Intelligence (GOFAI), a term coined by Haugeland (1985). GOFAI approaches were essentially based on the idea that cognition is computational in a classical sense. That is, in analogy to a digital computer, the brain can be seen as a central processing unit (CPU) that engages in symbol manipulation. In this view cognition is the act of processing information encoded in a symbolic fashion.

As such, AI systems and robots were endowed with control architectures based on 'sense-plan-act' loops that were executed in serial fashion: 1) perception of the environment through sensors, 2) building of an internal representation of the current situation based on perceived sensory values, 3) calculation of next action based on current situation and predefined goals, 4) execution of the action, and 5) back to step 1). A typical example of this type of system is the robot 'Shakey'[6] (Nilsson, 1984), which was one of the first in its kind, being able to reason about its environment and act accordingly. Such robots performed reasonably well in a static and predictable environment in which everything was controlled (e.g. a laboratory setting or a factory). Their performance was significantly less effective in unpredictable

[5]Cybernetics was defined by Norbert Wiener as "control and communication in the animal and the machine" (Wiener, 1948) and as such was about a theoretical understanding of a broad variety of feedback systems; including, but not limited to, AI systems.

[6]This robot was nicknamed 'Shakey' because of its peculiar way of moving.

dynamically changing environments such as an office with people coming and going, or an outdoor setting with changing weather and lighting conditions, background movement, uneven terrain etc. At the time however, most of the difficulties and limitations were perceived as resulting from a lack of computational power that would eventually be overcome through the development of more powerful computers (cf. Moore's law).

Within GOFAI architectures it was assumed that cognitive tasks could be decomposed in smaller subtasks (sense, plan and act) which could be solved independently in a heuristic fashion. So in order to solve a rather complex task, all a robot needed to do was to decompose it into manageable subtasks, find solutions for these and 'glue' it all together. The decomposition into subtasks however, turned out to be not trivial or even impossible on occasion. Also, issues with the subtasks themselves were problematic. For instance, how could sensory information be translated into a proper symbolic representation, how could discrete symbols be translated into continuous motor actions and, to anticipate a changing environment, how could planning be done in real-time while this tended to require a lot of computational time. Also the *Frame Problem* (McCarthy and Hayes, 1969), the problem of deciding which aspects of the environment are worth paying attention to (i.e. spending computational resources on) without actually computing them, plagued this approach[7]. Despite these problems GOFAI dominated AI research for decades.

New themes and ideas emerged in the 1980s. For example, Braitenberg (1986) showed how relative 'intelligent' behaviour could be achieved in very simple 'vehicles': software or hardware agents that exhibited certain reflexive behaviours like being drawn to light, avoiding light, approaching a target, moving in a certain pattern etc. The 'brain' of these vehicles was very simple, it consisted of clever coupling between sensors and actuators in such a way that when placed in the right environment, the vehicle displayed recognisable behaviours. For instance, the behaviour

[7]The Frame Problem is in fact not one specific problem, but rather a host of problems that all revolve around the notion that it is problematic for AI systems to make decisions about what should be computed, what is relevant and what can be ignored, without spending too much resources (time and computational power). For example, changes in the world that are not relevant for a certain problem should be ignored, but the only way to know if they are relevant or not is to examine them, which comes at a cost. Different variation on this problem and varying solutions have been proposed, see Lormand (1990).

'drawn to light' was achieved by cross-connecting light sensors and actuators: a light sensor on one side of the vehicle detecting an increase of light would increase output of the actuator on the other side, thus causing the vehicle to move towards the light. Braitenberg's vehicles illustrated that no complex high level cognition was required for simple forms of behaviour and such were examples of 'minimal intelligence'.

Also Brooks (1986) challenged the underlying principles of GOFAI by proposing the 'subsumption architecture'. This control architecture for autonomous agents consisted of multiple layers of behaviour which operated in parallel. Sensors where directly connected to motor outputs to allow for fast low level behaviour. On top of this other layers of behaviour could be added incrementally which could influence lower behaviour but also incorporated higher order objectives. Through parallel processing of all behavioural layers, the robot controller was proposed to be more adaptive toward changing environments than robotic controllers originating from the GOFAI approaches. Behavioural layers were hierarchically structured, where the simplest layers at the bottom served as reactive behaviour (e.g. obstacle avoidance) and more advanced layers which incorporated the higher order objectives could inhibit or exhibit the output of lower layers and hence strive for more long term goals. Although the subsumption architecture was inspired by nature (low level behaviour as 'reflexes', high level behaviour as higher cognitive functions like planning), each layer was still designed and tested by human programmers. Since each layer could influence all pre-existing layers complexity accumulated rapidly, which made the design of complex behaviours more and more challenging as new layers were added.

Even though the subsumption architecture was not the ultimate answer to the problems of the GOFAI approach, the 'behaviour based robotics' approach sparked the development of a broader movement. This paradigm, known as 'Embodied Cognition' or 'Embodiment' (Pfeifer and Scheier, 2001; Wilson, 2002), advocated the notion that symbol manipulating rule based systems were not suitable to capture the most basic characteristics of intelligence. Instead, the alternative proposal was a lower level, behaviour based, reactive, embedded and embodied approach, in which the environment, the brain and the physical system could not be seen independently

but should be studied as a whole. In this view the morphology of any cognitive system heavily influences its abilities and is therefore of crucial importance.

The view was broadened and researchers started to focus on biologically inspired techniques like neural networks and evolutionary algorithms because such techniques provided new means for the development of adaptive control. Also methodological tools from dynamical systems theory served the newly formed paradigm well. Dynamical systems theory describes the development of typically complex systems over time, so applying this to robotics, the robot's control program, the robot's phenotype and the environment are all regarded as one dynamical system in which interactions are taking place continuously. As such, cognition cannot be understood properly in isolation, but the environment in which it operates, the body in which the cognitive mechanism is embedded and the sensors that provide a connection with the outside world all play their respective parts. Moreover, boundaries between these aspects may be fuzzy, as cognitive processes may happen through utilisation of the body (e.g. counting with fingers) or through cognitive 'offloading' onto the environment (e.g. physically rearranging of objects to separate them into classes, or the use of tools such as notepads; Clark, 2008). Currently, the notion of embodiment has become fairly mainstream in cognitive robotics (Pfeifer et al., 2007).

1.3 Learning and developmental robotics

Another prominent aspect of the 'new AI' as described above is the appreciation of the fact that it is practically impossible to provide AI systems with all the knowledge they might need. As the real world is so enormously complex it is virtually impossible to predict what kind of knowledge an AI system should have in order to be able to deal with everything that might be encountered. Therefore, rather than providing robots with such 'innate' knowledge, they should be adaptive to their environment and learn skills and knowledge on the fly. Possessed with such flexibility, a robot can tune itself to whichever environmental factors are relevant for its current situation. Hence, within new AI approaches, adaptivity became an important topic as well.

Inspiration for this has been drawn from developmental psychology, as children

display an enormous capacity to quickly expand their abilities in a relative short time frame from the moment they are born. Even though infants are genetically disposed to certain characteristics, e.g. organisation tendencies in brain development, many of their cognitive abilities gradually develop through interaction with their environment, including their bodies, their caregivers, other children and the world. Developmental psychology has been strongly influenced by the pioneering work of Piaget (Piaget and Cook, 1953), who was one of the first to describe the different developmental phases children go through. Piaget's theories however, viewed the learning of children very much as a one-way process, in which the child gradually absorbs knowledge from the world through individual exploration. In contrast, the theories of Vygotsky (1964), put more emphasis on social, cultural and linguistic aspects of child development, placing the child in an interactive environment through which it learns.

These views have since been applied to robotics, as it was recognised that insights and principles from developmental psychology could yield interesting new ways in which robots can learn, thus giving rise to the field of cognitive developmental robotics (Dautenhahn and Billard, 1999; Asada et al., 2001; Lungarella et al., 2003; Asada et al., 2009). In this approach, rather than endowing a robot with certain fixed capacities and then 'exposing' it to the real world, a robot is viewed as a naive learner that gradually goes through different stages in development through interaction with its environment (Sandini et al., 2004). As such, the robot engages in incremental cognitive development in which simple capacities are acquired first and newer, more complex capacities are built on top of this (Cangelosi et al., 2010).

1.4 Language

All topics discussed so far relate to language in one way or another. It has been argued that language is at the root of cognitive abilities that are unique for humans and that set us apart from other animals (Bickerton, 1995); the ability to use recursion in language being the most unique feat (Hauser et al., 2002). Moreover, it has been suggested that the language faculty is intimately linked with the ability

to use concepts (Bloom, 2000), making it hard to distinguish between learning a concept and learning the meaning of a word. A long standing debate is the notion of linguistic relativism, also known as the Sapir-Whorf hypothesis, which states that the thoughts we have (including concepts) are determined by the language we speak (Whorf and Carroll, 1956). The strong version claims that all thoughts and concepts we can entertain are *determined* by the particular language that we speak, while the weaker version states that language merely *influences* concepts we have and behaviour we display (Kay and Kempton, 1984). Regardless of which version is supported, there exists a strong relation between concepts and language.

Related to this is the question of how both syntactic and semantic aspects of language are learned. Regarding syntax, one school of thought endorses the view that the capacity to learn language must be innate; a position which is most famously advocated by Chomsky (1986), who argued that because children are not exposed to sufficient linguistic material that could enable them to learn all complexities of language (the poverty of the stimulus), they must possess an innate capacity for this (known as Universal Grammar). Others have argued against this. For instance, Tomasello (2003) proposed a usage-based account of language learning which does not require the existence of a special language faculty. Rather, children are able to learn a language because they utilise cognitive skills like statistical analysis of observed patterns that are not unique for language but for other cognitive abilities as well, combined with an interactive process in which they recognise others (parents, teachers, playmates) as intentional beings through imitation, attention sharing, goal sharing etc. As such, language acquisition is a highly interactive, constructive and social process (Tomasello, 1992).

The ideas above are mostly focussed on how syntactic aspects of language (like grammar) are acquired; closely related to this is how semantic knowledge, i.e. meaning, can be learned. The latter becomes particularly pronounced in AI models of language learning, as a naive style of meaning representation using explicit definitions suffers from the 'dictionary' problem, the problem that the meaning of words

is very often defined in circular fashion[8]. Related to this is the symbol grounding problem (Harnad, 1990), which asks the question of how abstract symbols (e.g. language) can acquire meaning. Proposed solutions suggest that low level meanings must be grounded in sensorimotor contingencies on which more abstract notions can be based (Massé et al., 2008). Alternative views have placed more emphasis on the interactive and social nature of language acquisition. For instance, Steels (1997) showed how symbols and meanings can become shared in a population of agents through a process of continuous social interaction with other agents. He proposed the framework of language games; a mechanism in which artificial agents, triggered by what they observe in their environment, engage in linguistic interactions through which they acquire word-meaning combinations and gradually align these with other agents in the population.

1.5 Human-Robot Interaction

Taking the embodiment thesis to heart, we are not only concerned with simulated agents, but also consider the learning agent as embodied in robotic hardware. This brings us to the topic of Human-Robot Interaction (HRI) and social HRI in specific. HRI as a research field has considerably developed in recent years. With the increasing availability of advanced robotic platforms that are meant to interact with people in unstructured, informal settings[9], questions about how exactly these robot systems should interact with people have become more articulated. Given that robots are becoming more and more sophisticated in their appearance and behaviours, insights gained from the more established field of Human-Computer Interaction (HCI) may

[8] A typical illustration of this problem is the following. Dictionary.com defines the word 'fact' as "something that actually exists". The word 'actual' is then defined as "existing in act or fact", the word 'existing' is defined as "to have actual being" and finally the word 'being' is defined as "the fact of existing". Thus, the circle is complete and nothing is learned about the meaning of these words, other than that they define each other in a circular fashion.

[9] This is in contrast with industrial robotic applications that typically operate in tightly controlled and highly predictable environments. For this type of robots there is a notion of HRI as well, as they are operated by people to some extent, but this is of a drastically different nature than the HRI we discuss here, as personnel operating industrial robots will typically have received formal training. In contrast, the robots we are concerned with are envisioned to operate in environments with the general public; thus people that come into contact with these robots are not expected to have received any specialised training.

not be directly transferable to an HRI context. The new discipline, HRI, has steadily been gaining ground in the last decade, building on foundations from HCI (e.g. anthropomorphism) but gearing this towards a view in which humans and robots are seen as social partners (Breazeal, 2004). The fact that people tend to treat robots as social entities opens up new possibilities for interaction (Duffy, 2003). More than an 'ordinary' computer, a robot can elicit social behaviour towards it and effectively utilise social channels that come naturally to people, both verbal and non-verbal (Breazeal, 2000).

This can be beneficial for learning as well. People display natural tendencies of tutoring behaviour towards others and in particularly towards infants and children; e.g. by providing isolated, structured learning examples which are easy to grasp. A robot could greatly benefit from these tendencies; the more a robot behaves like a human learner, the more effective the teaching might become. Important in this regard is the ability of a robot to utilise social channels in a correct manner. As we want to avoid the uncanny valley[10] (Mori, 1970), currently the aim for a robot is not to 'mimic' human behaviour in every aspect, but rather behave in such a way that people find the interaction natural. That is, a robot should behave according to people's expectation, and this is best achieved through transparency with respect to the robot's capabilities. For instance, the presence of physical ears on a robot head is an indication that it can hear, dexterous hand movements may indicate that the robot can understand people's gestures and so on. The robot should portray an image that is overall coherent. HRI will be more extensively described in section 7.1.

1.5.1 Social learning

An important aspect of human development is that it is strongly embedded within social context (Rogoff, 1990). Indeed, virtually all important developmental phases an infant goes through happen within a social environment, which is typically shaped

[10]The uncanny valley is the observation that the more human-like robots become, the more familiar people might feel towards it. However, at some point a robot might reach a resemblance that is very close to human, while at the same time it is obvious that it is still an artefact. This state of 'almost like human but not quite' can feel uncanny and eerie to people, causing a drop in their familiarity towards the robot.

by the infant's caregivers (Gauvain, 2001). Language acquisition and the development of a capacity for concept use are no exception to this (Tomasello, 1999; Akhtar and Tomasello, 2000). As such, a theory seeking to explain concepts from a developmental perspective will need to acknowledge this social component.

Consequently, the notion that social context is also relevant for knowledge representation and learning in artificial system has gained more attention in recent years. Indeed, social aspects in robotics are recognised as important, see for instance Fong et al. (2003). Also pioneering work of Breazeal (2000) contributed to the consensus that specifically robots, in contrast to artificial systems in general (which might be embodied in robots but might also exist in e.g. simulations), are easily viewed as social partners because they can tap into social channels that are perceived as natural by people. As such, a learning robot can benefit from learning strategies rooted in social interaction, by manifesting the appropriate social behaviours. Social robots that are able to effectively utilise these social channels might be particularly suited to learn from humans in an ongoing, interactive fashion.

A relative new view on machine learning and artificial intelligence is the so called socially-guided machine learning (SG-ML) as advocated by Thomaz (2006). In the SG-ML paradigm classical machine learning techniques are augmented with social guidance in order to make the learning more effective. This proves particularly useful in the context of cognitive robotics, as the context in which these robots tend to operate may very often naturally provide the required guidance. In other words, no 'artificial' or 'contrived' scenarios need to be employed, as a social environment is naturally rich in learning opportunities and social partners are generally available to provide guidance.

Recently, such interactive machine learning has gained more attention; different studies have demonstrated that robots can benefit from employing interactive strategies in which the robot learner is not passively absorbing new knowledge, but rather actively engages in the learning experience through social interaction with a human teacher (Brooks et al., 2004; Thomaz and Breazeal, 2008). Also different robot behaviours were investigated by Cakmak et al. (2010), indicating that robots

may benefit from active querying as opposed to standard supervised learning. Human tutors appear to appreciate an active learner, but like to stay in control at the same time. Thus, balanced behaviour may be most optimal, and this can vary for different users. Moreover, optimal robot behaviour might require fine-grained understanding of the social situation in order to be effective (Knox et al., 2012); as such, more exploration of appropriate robot behaviour remains to be done.

1.6 The thesis

Within this thesis the various themes as described in the previous sections come together. Concepts are important for human cognition and therefore AI systems should be able to learn and use them. Recent trends in AI, drawing inspiration from developmental psychology, have focussed on embodiment, adaptivity and learning with promising results. It is believed that language, as a bearer of conceptual knowledge, can be acquired in a fluid constructive process through interaction with the environment. As such, a developmental approach to concept learning can be promising. In this view, an AI system, embodied in a robot, gradually acquires concepts through social interaction with human tutors. Moreover, the manner in which this happens is one of active learning, in which the learner actively tries to influence the learning experience though utilisation of social cues, as to obtain better results.

Having set the stage, the following general questions can be formulated:

- *How can an artificial system be endowed with a conceptual model that bears resemblance to human concept learning and how might social interaction impact on concept acquisition through such a model?*

- *Given variations in embodiment and specifically perception, how might social concept acquisition overcome these differences?*

- *Can an artificial system equipped with such a conceptual model learn through social interaction with humans?*

These questions are addressed as follows. First, various theories about concepts are explored from two perspectives, one of these focussing on psychological theories of how human use, learn and adapt concepts as part of their cognitive development (the human perspective, section 2.3), and the other relating notions of concepts to machine learning theories and techniques (the machine perspective, section 2.4). Combining these two perspectives, it then is described which aspects of concepts are deemed important for a model of concept learning applied to AI, addressing important notions from both human and machine concept use (section 2.5). This results in the formulation of a computational model that combines two frameworks: *Conceptual Spaces* (Gärdenfors, 2000a) as a means of representing concepts and *Language Games* (Steels, 1998) as a means of social acquisition of concepts. This computational model is described in detail in chapter 3.

Using this computational model of concept learning as a starting point, some aspects regarding its basic functionality are explored. Specifically, three different aspects are addressed: 1) two forms of concept learning and testing are compared, one based on classical language games, and the other as a form of more direct learning (section 4.2); 2) different means of representing concepts in a conceptual space are compared (one based on Discrimination Games, one based on k-Means and one based on Self-organising maps, section 4.3); and 3) the model's ability to represent animal concepts in a manner compatible with prototype theory is tested (section 4.4).

Having established the computational model and its basic functionality, we subsequently move to the more experimental part of the thesis, in which the notion of active learning, embodiment and social scaffolding of concept learning are investigated, both using simulations and a real-world setup that consists of people interacting with a novel robotic face (section 7.2). Specifically, the following aspects are addressed experimentally. To investigate the effectiveness of active learning in a language acquisition setting, agents are endowed with means to actively influence their learning experience. This allows them to more effectively learn concepts through interaction with other agents (chapter 5). Next, we explore the effect of dif-

ferent embodiments on this type of learning (chapter 6). Both simulated agents and robots with differences in their perceptual capabilities engage in learning interaction; the flexibility of the learning dynamics allows them to overcome their perceptual differences (section 6.2). These experiments are then extended to simulation that more accurately models human vision, allowing us to make the argument that these learning mechanisms might account for the phenomenon that people's subjective colour experience hardly seems to differ, despite large physiological differences in the retina (section 6.3). The next steps include extension of the findings into a robotic setup (chapter 7). First the ability of people to read a novel robot's gaze is investigated, as mutual gaze understanding contributes to HRI (section 7.3). Having established that the robot can be an effective social partner, experiments are presented in which the robot learner employs the active strategy that was explored in simulation, by utilising social cues, through interaction with a human teacher (section 7.4). Results indicate that active learning gain also holds in a real HRI setting.

1.7 Contributions

In light of the description of the thesis provided above, the original contributions of this work can be summarised as follows:

- A demonstration is provided of how two frameworks (conceptual spaces and language games) can be combined into a functional model which is able to learn concepts in a manner that is consistent with the prototype/exemplar theory of concepts and which places emphasis on the social and interactive nature of concept learning; both as a formalisation and through a working implementation.

- The language game framework is augmented with interactive features to acknowledge the interactive nature of learning. In a series of experiments it is shown that active learning, in which the search space of an agent is more effectively explored, constitutes an improvement over the classical approach.

- Despite physiological differences people display a remarkable agreement with respect to colour terms. Through a computational approach which models the development of colour terms as interactions between agents with perceptual differences, an alternative explanation of how this phenomenon can come about is provided.

- Mutual gaze understanding is important for joint attention and learning, and as such robots employed in HRI need to support this. Through experiments in which human participants read the robot's gaze, the LightHead robot is established as a novel HRI platform.

- Combining the aspects described above, it is experimentally established that active learning aspects in human-robot tutelage also hold when employed in an actual robot setup in which human participants teach concepts to the LightHead robot.

1.8 Structure

The structure of the thesis is provided below, with a description of the content for each chapter.

- In this introductory *chapter 1* the main themes that are important for this thesis have been introduced, along with the thesis narrative, the contributions and the structure.

- In *chapter 2* a more extensive background on concepts is provided. The concept of concept is discussed, and theories about concepts are described from two different perspectives. The human perspective deals with theories from psychology about how concepts are used by humans, and the machine perspective describes theories that are more based on computational modelling of concepts and how this is done from an AI perspective. These two perspectives are then related to each other and it is described how those aspects of concept

learning that are deemed important are combined into a functional model that supports social concept learning.

- *Chapter 3* introduces the computational model that is used for the experiments described in this thesis. The model is based on the combination of two frameworks: conceptual spaces and language games. We explain how this combination leads to a functional model that fits in with the background provided in chapter 2 and provide some of the more technical descriptions of the model, including evaluation methods and a description of relevant parameters.

- In *chapter 4* some of the basic functioning of the model is explored through simulations. A baseline simulation is provided, against which we compare direct instruction as an alternative to language game learning. Furthermore, the partitioning of the conceptual space as resulting from the discrimination game is compared to two alternative classification methods: k-Means and self-organising maps. Finally the model's abilities to display prototype and typicality effects are explored.

- *Chapter 5* concerns a series of experiments centred around the notion of interactivity in learning. It has been fairly well established that learning is an interactive process in which both teacher and learning actively contribute. Using colour as an exemplar case, augmentations of the language games framework are considered that allow an agent to be more actively engaged in the learning process. The effect of these augmentations on the learning process is tested in an experimental setting; particularly the notion of active learning turns out to be useful.

- The effect of embodiment is explored in *chapter 6*. A series of experiments is run in which artificial agents and robots are endowed with different perceptual abilities. The effects these perceptual differences have on the learning and alignment of colour terms are discussed. In a second series of experiments we argue that the computational models can provide an alternative explanation for the phenomenon that people hardly experience any difference in colour

perception, despite striking differences in the physiology of the human retina.

- In *chapter 7* robotic aspects are included. First the LightHead robot is introduced, which constitutes a novel robotic platform for use in HRI studies. As gaze is very important for mutual understanding in HRI, the ability of participants to read the robot's gaze is tested in experimental fashion. Next, experiments in which human tutors teach the robot concepts are described, showing that the active learning principles described in chapter 5 also apply in an HRI setting.

- *Chapter 8* provides an overview of the work covered in the thesis, reflects on aspects that are related to the thesis in a broader sense and discusses some topics of future research.

Chapter 2

Concepts

This chapter sketches a theoretical background of concepts. First, a brief historical overview of thinking about concepts throughout the history of science is provided and relevant theories of conceptual modelling are discussed. Theories about concepts and classification are explored from two different viewpoints: one is the human perspective, in which we describe theories related to human concept use, and the other is the machine perspective, in which some machine learning techniques related to concept learning and classification are discussed. Then, we discuss learning in a broader sense, with an emphasis on the social aspects of concept learning. Finally, we describe how insights from both perspectives, with the inclusion of a social component, can be combined into a computational model which is subsequently is used for the experiments reported in this work.

2.1 A note on defining *concept*

Given the prominent role of concepts in this thesis, it would be fairly common to start with a definition of what exactly a concept is understood to be; or in this case, what the *concept* of concept might be. However, attempts at providing such a definition might prove to be somewhat futile, predominantly because no consensus regarding a definition of concepts can be found in the literature. As such, rather than providing an exhaustive list of all definitions that have been proposed, or trying to formulate a new one, we will discuss some of the issues related to formulating a

definition of concepts and suggest that perhaps a strict definition is not necessary.

Within the literature the terms *concept*, *category* and *class* are frequently used interchangeably, as are *conceptualisation*[1], *categorisation*, *classification* and *discrimination*. In case of the latter, all terms refer more or less to an act of assigning some kind of data as belonging to some kind of group. It is sometimes understood that a category refers to a kind of ordering, organisation or grouping of things in the world, and that a concept is an internal representation of an agent[2] for this particular category. This distinction between categories and concepts can lead to ontological questions such as the following: do categories exist in the world without an agent to make this categorisation, and if they do not, how could they be distinguished from concepts? The causal relation between concepts and categories is also unclear if this distinction is assumed: do people form concepts in response to the categories they perceive, or do they form concepts as a means of organising perception and then impose these conceptual structures on the observations they make? (Goldstone and Kersten, 2003).

In (Davidsson, 1996, p. 62) several definitions of concept and category are discussed; the conclusion is that the terms concept and category do not have a clear definition and need to be explicitly defined within the theoretical framework in which they are used. However, as Murphy (2002, p.5) notes, "being too fussy about saying *concept* and *category* leads to long-winded or repetitious prose (...) with little advantage in clarity". Thus, one can question the need to explicitly define the difference between a concept and a category. Instead, we will follow Murphy's example and loosely use the terms *concept* and *category* as synonyms unless the difference is relevant; if this is the case the difference will be made explicit.

Sometimes a difference between perceptual categories and conceptual categories is postulated (Mandler, 1997). This proposal usually goes along the following lines. Perceptual categories are those based on observable features, such as colour, shape,

[1] *Conceptualisation* can refer to the formation of concepts, i.e. the creation of mental objects that are concepts, but is also commonly understood as the practice of making new ideas concrete, e.g. concept art as a visualisation of some design or idea. However, in this text the first reading is meant.

[2] Agent in this context is to be understood in a broad sense: an entity that perceives and acts within an environment. This includes natural agents such as animals and humans as well as artificial agents (both in software and hardware).

size, texture etc. Conceptual categories on the other hand rely on distinctions that may or may not be observable, e.g. properties such as "has blood" or "is alive". The latter tend to be more abstract and are typically learned later in a persons lifetime. Whether or not there is a real distinction between the mechanisms that give rise to these supposedly different types of concepts is still debated in the literature (see Quinn and Eimas, 2000; Rakison and Oakes, 2003). Therefore, we will not presuppose a difference between perceptual and conceptual categories. Rather, all categories or concepts are considered equal in terms of their structure and underlying mechanisms of acquisition. Some might be very concrete with easily identifiable properties, e.g. DOG, CAR and APPLE, while others are more abstract (conceptual), e.g. JEALOUSY, INNOVATION and PEACE.

2.2 Theories concerning concepts and classification

Since ancient times concepts have been considered as fundamental to human knowledge and reasoning. As such, concepts are commonly viewed as being at the heart of cognition, i.e. as building blocks from which higher level cognitive processes can come about. Throughout the history of science mankind has been formulating theories about how concepts are formed and how they constitute human cognition. However, in many cognitive theories there is not always a clear distinction between concepts and other aspects of cognition such as knowledge, memory or semantics. Hence, cognitive theories may use different terminology but may nevertheless be concerned with the same or similar processes, structures and organisational principles of cognition.

2.2.1 Different perspectives

The disciplines involved in the formulation of models about concepts and categories are vast; they include philosophy, psychology, neuroscience, linguistics, biology, computer science and artificial intelligence just to name a few. Within all these fields

theories about concepts and classification have been developed, albeit using different terminologies. In an attempt to provide some insights in this broad theoretical span, some relevant theories are discussed from two different perspectives: the *human* perspective and the *machine* perspective[3]. Applying the human perspective and the machine perspective to artificial intelligence, these two perspectives highlight the two different goals of AI (as briefly discussed in chapter 1).

On the one hand, AI is about understanding human intelligence[4] through the creation of models of cognition. Through application of synthetic means it is possible to systematically explore a wide range of different models with different parameters; all of which might be relevant for cognitive functioning, but which may be hard or even impossible to manipulate in living organisms because of various practical or ethical considerations. The goal is to gain understanding of natural intelligence through artificially creating it.

On the other hand, undertakings in AI can be aimed at the creation of machines that have some kind of intelligence and are capable of performing in a dynamic, ever changing world populated with other intelligent agents (human or artificial). Thus, the goal of this endeavour is to create some kind of machinery that is able to perform its tasks in an intelligent manner. To reach this goal an understanding of human intelligence might be required, as humans prove to be capable of very versatile, adaptive and creative behaviours in a wide variety of circumstances and as such constitute a 'solution' to the 'problem' of intelligent behaviour. Hence, from the machine perspective this understanding of human intelligence is not a goal in itself, but merely an exercise in 'borrowing' of ideas and serving as a source of inspiration[5].

Thus, where research from the human perspective is primarily concerned with the

[3]One might argue that humans can be seen as biological machines; as such a *machine* perspective in contrast to a *human* perspective becomes meaningless. However, here we mean to highlight the distinction between humans (or other organisms) as vehicles for natural intelligence and machines (artefacts) as vehicles for artificial intelligence.

[4]Or rather natural intelligence. Human intelligence might be considered as the pinnacle of natural intelligence, but other forms of animal intelligence are not of lesser interest. Indeed, the study and understanding of lower forms of intelligence might provide a prerequisite for understanding higher forms.

[5]Of course, in day-to-day execution of AI research, the two perspectives can blend seamlessly; a lot of researchers tend to have both an interest in natural intelligence and a desire to build machines that can display intelligent behaviour.

understanding of natural cognition and employs computational models to this end, the goal of research from the machine perspective is to apply these computational models to artificial systems in order to achieve intelligent behaviour. In the first perspective psychological, biological and/or neuroscientific plausibility is of major importance, as this is at the heart of the problem under investigation. A good model of human intelligence will need to be consistent with empirical data from these disciplines. In contrast, from the machine perspective the ultimate goal is to develop a system that just works; therefore biological plausibility could be compromised if it turns out that better results can be achieved by taking certain 'short-cuts' that may be not biologically plausible. However, it could turn out to be that in order to achieve truly functional artificial cognition such short-cuts may not be available. That is, the only way to build a truly intelligent system is to do it in the same way as nature has done. Whether or not this is the case remains to be seen; if anything, biological intelligent systems provide at least one working example of an 'implementation' of intelligence.

Having described the general idea of studying concepts and categorisation from the human and machine perspective, we will now for each of these perspectives discuss some relevant theories in more detail.

2.3 The human perspective

2.3.1 The relation between concepts and words

The difference between knowing a word and knowing a concept is often quite hard to describe. Indeed, in a lot of literature this is (implicitly) regarded as one and the same process, as a person knowing the meaning of a certain word is the *de facto* proof that this person entertains a certain concept. And the other way around, if one has a concept of something, one usually knows the word for this. Yet, intuitively it seems conceivable to have a concept for something without knowing a word for it, i.e. having a sense of something without knowing the precise words to describe this. Also various studies have shown that prelinguistic infants are able to form concepts

before they learn to speak, e.g. Roberts (1988); Quinn (2003); Rakison (2003). This suggest that words and concepts are at least partially independent.

Related to this is the Sapir-Whorf hypothesis which states that language influences thought to various degrees. As briefly mentioned in section 1.4, the strong version, also known as linguistic determinism, claims that language determines thought; a weaker version, known as linguistic relativism, states that language merely influences thought. This notion gained renewed attention as a series of psychological experiments demonstrated how perception of stimuli and use of categories is influenced by the words we know; this has been notably demonstrated for categories of time, colour and space, e.g. (Gilbert et al., 2006; Majid et al., 2004; Roberson et al., 2008). In addition, field studies such as Gordon (2004) have demonstrated that the lack of certain words can have severe impact on cognitive abilities such as counting; not having words to describe quantities greater than two[6] makes it difficult to distinguish between different sets of quantities.

From a developmental perspective it has been shown that young children, in addition to learning directly from sensory exploration, rely on linguistic labels to acquire the meaning of words. Xu (2002), for example, demonstrated how linguistic labels help 9-month old infants to establish a representation for different objects; learning without linguistic labels, or with the presence of tones, sounds or emotional expressions is not effective. This implies that language is crucial in acquiring novel concepts from a very early age on. Plunkett et al. (2008) came to the same conclusion in a tightly controlled experiment where they demonstrated how category formation in 10-month old infants is influenced by linguistic labels. Linguistic labels also have an effect on category learning in adults; adults who learn a new category did so significantly faster and showed more robust category recall when the learning experience was accompanied by novel linguistic labels (Lupyan, 2006; Lupyan et al., 2007).

While language clearly influences cognition and the learning and use of concepts, other studies have suggested a relative independence between words and concepts. For instance, Dufour and Kroll (1995) described experiments with bilingual people,

[6]Every quantity greater than two that needs to be named is referred to as "many".

which indicated that words in different languages can share a same underlying conceptual structure. Additional evidence comes from Goldin-Meadow et al. (2005), who studied deaf children (both American and Chinese) that were not exposed to adult language models; neither acoustic nor through sign as deaf children from non-deaf parents are typically only exposed to conventional sign language when they are adolescents. Nevertheless these children frequently invent gesture systems called 'home sign', to be able to communicate with their surroundings (Goldin-Meadow, 2003). It was observed that such children develop the ability to express generic utterances ("birds fly") in sign language as well. The use of such generic utterances (and also the presence of a bias to do this more frequent for animal categories than for artefacts categories) suggested that these children had developed conceptual structures that are similar to non-deaf children who grow up in 'normal' circumstances in which linguistic input is typically abundant.

In summary, there exist ample evidence suggesting that language influences cognition, perception and category interpretation; while in addition there are also indications that conceptual structures can be formed without language and can operate relatively independent from it. As such, within this work a standpoint is adopted in which words and concepts are treated as separate notions (that is, agents will have a lexicon containing word labels and a different structure to represent concepts), albeit with a strong influential connection between them. Indeed, as will be seen in work reported later on, linguistic descriptions and interactions serve as guiding mechanisms for the formation of concepts.

2.3.2 Ancient theories about concepts

The best known ancient scholar who wrote about a theory of concepts is the Greek philosopher Plato, who described the nature of reality in his *theory of Forms*. As described in Kraut (2011), according to Plato, every worldly observation people make about some object is just a flawed reflection of this object's true form. These true forms are entities that reside in a realm that is different from the world we can perceive with the ordinary senses. This realm is more 'real' and perfect, and is

populated with entities that are the perfect versions of what we might perceive in the normal world. These entities can be concrete like "table", "bed" and "horse" or more abstract entities like "goodness", "beauty" and "just". According to Plato, analogue to prisoners in a cave watching shadows of real things (Plato's famous Allegory of the Cave), everything we see and hear around us is but a bleak shadow or echo of its true form. These forms are said to reside in the non-material realm and constitute the only true knowledge. Ordinarily people have only access to the imperfect copies of the forms through ordinary senses like sight, smell, touch, etc. However, the true forms can be reached through disciplined exercise of the mind, through reason and philosophy.

The other well known Greek philosopher, Aristotle, had a rather different view on concepts and categories. As opposed to Plato, Aristotle did not subscribe to a realm in which the perfect versions of all things in the world resided. Rather, things in the world could be classified according to ten basic categories, such as 'Substance', 'Quality', 'Quantity', 'Place', 'Time' etc. These categories are both presumed to be both "exhaustive and irreducible" (Shields, 2012), so that everything in the world can be classified according to this.

Another prominent philosopher who spoke about concepts was John Locke (as described in Uzgalis, 2010). According to Locke there are *general ideas* which can be simple or complex, where a combination of simple ideas can create a complex idea. In opposition to Plato, these ideas are purely internal and come into existence through the human senses, rather than residing in some kind of abstract realm.

2.3.3 The classical theory of concepts

Rooted in ancient philosophy (e.g. Aristotle's categories), a definition style of categorising observations in the world remained dominant even within substantial periods of modern psychology. This view, known as the 'classical theory of concepts', assumes that concepts, and specifically lexical concepts (i.e. meanings of a word) are based on definitions. Thus, the concept of CAR might be defined in terms of descriptive properties such as 'means of transportation', 'moving construction', 'having

	moves	flies	eats	lives	has wings	has wheels	carries people
AIRCRAFT	x	x	-	-	x	x	x
BIRD	x	x	x	x	x	-	-
CAR	x	-	-	-	-	x	x

Table 2.1: Example of the definitional structure of concepts under the classical theory. In this example AIRCRAFT is defined as something that moves, flies, has wings, has wheels and carries people; consequently, everything that has these properties must be an AIRCRAFT.

four wheels', 'having a steering wheel', 'carrying passengers' etc. A definition of a concept is viewed as a list of necessary and sufficient properties that describe the concept. In order for something to be called a particular concept, it must have all the necessary properties; and vice versa, if some object has these sufficient properties, it is an instance of this concept. For example, BACHELOR may be comprised of the properties 'adult', 'male' and 'is not married'; consequently, everything that is an adult male and is unmarried is therefore a BACHELOR (Murphy, 2002; Margolis and Laurence, 1999; Smith and Medin, 1981). Table 2.1 provides an illustration of such definitional structure. Moreover, every object is viewed as belonging to some category, there is no space for in-between cases. As such logical statements like "x is a CONCEPT" are either true or false. Also, all members of a concept are considered equally good instances of this particular concept.

2.3.4 Problems with the classical theory

For a long time it was accepted that concepts could be defined in terms of these necessary and sufficient properties. But this classical view turned out to have some fundamental flaws, which are discussed below. Alternative theories were proposed, many of which can be viewed as a response to the classical theory in one way or another.

Two kinds of problems caused the downfall of the classical theory. The first is theoretical in nature. That is, the classical theory specifies that all concepts should have a neat logical definition, but it turns out that for a lot of concepts it is very hard to come up with such a definition. Despite trying hard for decades, for a lot of concepts it appears to be simply impossible to come up with a list of

defining features. Wittgenstein identified this problem in his classic investigation of the concept GAME (Wittgenstein, 1953). He tried to find a common property that all games should have, but this turned out to be non-existent. Wittgenstein solved this by proposing a structure of 'family resemblance', in which all games have some overlapping properties and as such resemble one another, but there is no ultimate defining 'game' property that all games posses.

The second type of problems for the classical theory is more empirical in nature. Even if logical definitions for concepts can be found, in reality there are a lot of things in the world that fall outside of these defining boundaries. For instance, carpet may be considered as belonging to FURNITURE by some people but not by others, and what about a waste bin? They seem to reside on the boundaries of what is and what is not counted as FURNITURE. Also, people may be unaware of certain definitions and assign objects to a particular concept, even though this is not consistent with this concept's definition. For example, many people regard a tomato as a VEGETABLE even though it scientifically classified as a FRUIT; they may count a dolphin as a FISH and not as a MAMMAL and so on. It turns out that in every day situations people do not seem to use very strict definitions with respect to the classification of objects. Furthermore, people do not consider all members of a concept as equal, but instead think of some members as better examples of a concept than others.

These two types of problems might be summarised as follows: 1) definitions are very hard to find for a lot of concepts, and 2) even if definitions exist, in practice people do not appear to use them consistently. As such, these problems were considered as major flaws in the classical theory of concepts and alternatives seeking to remedy these problems were proposed.

2.3.5 Prototype theory, exemplar theory and similarity

The observation by Rosch (1973) that many everyday concepts are *prototypical* in nature constituted a major problem for the classical theory. Rosch showed that many concepts cannot be logical defined (as was common practice under the classical

theory) because they show *typicality*, i.e. humans judge certain instances of a specific concept to be more typical than others. For example, for the concept BIRD, a robin is thought to be more 'bird-like' than a penguin and a banana is more typical for FRUIT than a pomegranate[7]. Hence, it seems that instances of a concept exhibit a graded membership to an idealised prototype, so that some instances are judged to be more typical for this concept than others. This cannot be explained by the classical theory, as any instance that falls within the definition is considered an equal member of this concept.

The prototype theory has been around for quite a while, with many different flavours (Posner and Keele, 1968; Reed, 1972). The general notion is that concepts are represented as some kind of idealised version constructed from examples that people have experienced throughout their life. So, for the concept BIRD people have a prototype that represents the ideal bird, and any encounters they have in the real world is matched with this prototypical version. The more similar an observation is to the prototype, the more likely they are inclined to consider the observation as an instance of this concept. Given the wide variety of birds, it seems unlikely that all members of BIRD could be represented by one single prototypical bird that actually exist (Murphy, 2002, p. 42). So a prototype should be thought of as a summary representation that specifies the properties of the concept, where some properties are more important than others. Not all properties are necessary, as in the classical theory, but rather they describe which properties instances of the concept in general tend to posses. The process of identifying an object in the world entails a matching to known prototypes. This matching takes the form of a similarity measurement, rather than a logical "does it ticks the boxes?" type of analysis. For this similarity measurement several methods have been proposed, this is discussed in more detail in section 3.1.

The prototype theory also has its problems, for instance the problem of compositionally (i.e. combining simple concepts into more complex ones; Osherson and Smith, 1981), but solutions have been proposed for this as well (Smith and Osherson, 1984; Smith et al., 1988).

[7]Of course, this may be culturally different.

Exemplar theory The exemplar theory as proposed by Medin and Schaffer (1978) is historically the main competitor for the prototype theory. In the exemplar view concepts are not seen as idealised prototypes that exhibit the concepts' most prominent features, but rather a concept is comprised of the collection of all exemplars a person has encountered in life. So, the concept of CAT consists of the collection of all cats one has experienced. Upon perceiving something new in the world, people mentally compare this to things they have experienced previously. If the newly perceived object bears considerable resemblance to known objects that happen to be of the same concept, then most likely the new object is also an instance of this concept. As with the prototype theory, this entails a similarity measurement. The main difference is that concepts are not stored as summaries, but all individual exemplars are stored in memory and together constitute a concept (Nosofsky, 1986).

Prototype and exemplar theory The debate whether or not prototype or exemplar theory better explains empirical data has quite a history, with some researchers advocating models based on prototype theory (Smith et al., 1988), others favouring models based on exemplar theory (Nosofsky and Zaki, 2002) or some arguing for hybrid models (Voorspoels et al., 2011). These days the debate seems to have subdued somewhat, although researchers may still embrace one theory or the other.

Both prototype and exemplar theory are based on the notion of some kind of space in which concepts reside, and which allows for comparison in term of similarity through distance measurement. This has been established as a fundamental property for conceptual modelling and has been at the basis for many theories that followed. Different accounts exist of how this similarity matching is modelled best (Hampton, 1995).

2.3.6 Other accounts

Some alternative accounts of conceptual modelling that do not subscribe to the notion of similarity have been proposed. Two of these frequently feature in the conceptual literature and are discussed briefly for the sake of completeness. However, they are not further considered in the rest of this thesis.

Theory-theory The theory-theory or knowledge approach (Murphy and Medin, 1985) regards concepts as a part of mental theories and general knowledge a person holds about the world, in a manner similar to scientific theories. As such, concepts do not exist as isolated chunks of knowledge, but learning a new concept always happens in interaction with the general knowledge a person has and entails the incorporation of this new information with already existing knowledge. Combined with everything else a person knows, the new concept forms an intricate web of knowledge. This framework of knowledge needs to be consistent; learning something new may have implications on what is already known and likewise, what is already known may influence how a person learns something new. The development of consistent sets of concepts within this framework may bear similarities to how scientific theories unfold. Throughout a persons lifetime concepts are enriched through new experiences; as such the sets of concepts gradually evolve and at some point may become incommensurable with other sets of concepts (Carey, 1991). Such a paradigm shift may bear similarities to a scientific revolution as described by Kuhn (1962).

Conceptual atomism Conceptual atomism as a theory of concepts (Fodor, 1981, 1998) is a somewhat radical position, as it advocates that concepts do not have a structure that specifies their properties, but rather have no structure at all. In this view all concepts are seen as atomic primitives. Conceptual atomism is mostly a reaction to the fact that previous theories had problems with a seemingly important aspect of concepts: the fact that many concepts are compositional. One of the main advantages of conceptual atomism is that it can handle compositionally rather well, something the theories mentioned afore tend to struggle with. Due to the atomic nature of concepts they can very easily be applied in symbolic reasoning, including compositional constructs. A main drawback is that it views concepts as innate, and thus supposes that also obvious artefacts like MICROWAVE or VOLLEYBALL are somehow innate. How acquisition of new concepts is supposed to happen is generally underdeveloped.

2.4 The machine perspective

The machine perspective on concepts and classification uses cognitive theory to create 'smarter' and 'more intelligent' machines. For a lot of non-trivial cognitive challenges (i.e. things people do, like navigating in an unknown environment, abstraction, reasoning and speech production and recognition) it has become more and more apparent that relatively simple, algorithmic solutions (GOFAI) may not be able to provide solutions. Rather, a deeper understanding of true human cognition seems necessary. A description of this development from GOFAI to new AI has been given in section 1.2.1; this section is mostly concerned with the so called new AI methods.

From the machine perspective modelling of concepts tends to be less articulated. A lot of research has focussed on the problem of machine learning, that is, how an artificial system can acquire new knowledge in automated fashion. For instance, a typical machine learning problem is classification: to assign a class or category to perceived stimuli. There are countless approaches solve this problem, which in general involve some kind of labelling of examples by a teacher while the system learns the right abstractions in the case of supervised learning; or, in the case of unsupervised learning, the system autonomously discovering the input space and developing some kind of representation for this. These types of machine learning problems tend not to be regarded as concept learning *per sé*, but the process of appropriately classifying observations bears considerable similarities to having conceptual structures and applying these effectively. Thus, there exist considerable overlap with the human perspective, but methods and terminology tend to be different.

Because of the vast body of literature that exist with respect to machine learning, the aim in this section is not to provide a general overview of all machine learning techniques that provide means of classification. Rather, we will link some common machine learning techniques to the theories of conceptual modelling described in the human perspective above. Furthermore, we provide a synthesis of how a combination of different approaches can lead to a computational model which can be employed within social concept learning.

2.4.1 Machine Learning and Classification

Machine learning is commonly viewed as the means by which an artificial system (a computer or a robot) can self-improve or acquire knowledge in an automated fashion. A classical topic within machine learning is classification, the problem of finding classes for a set of datapoints based on their features and generalising over these to classify new data. Classification as such bears similarities to concept use in humans, as the objective is to learn/identify classes (concepts) and to interpret new observations in light of what is already known. Numerous classification algorithms have been developed over the years, e.g. see Mitchell (1997) and Kotsiantis et al. (2007) for a good overview of various machine learning techniques and an overview of some well known supervised algorithms. One of the most commonly used classification algorithm is k-Nearest Neighbour (see e.g. Duda and Hart (1973) for a description), in which an unknown data point is assigned as belonging to the same class as the majority of its k neighbours. This requires knowing the classes of the surrounding data points and as such is a form of supervised learning. Another technique that operates in an unsupervised manner is k-Means clustering. k-Means is a commonly used clustering algorithm for solving the k-Means problem that can be defined as follows: (definition taken from Kanungo et al., 2002) given a set of n data points in real d-dimensional space, $\mathbf{R}^d$, and an integer k, the problem is to determine a set of k points in $\mathbf{R}^d$, called centers, so as to minimize the mean squared distance from each data point to its nearest centre. The k-Means algorithm (also known as Lloyd's algorithm; Lloyd, 1982) groups input data into k classes in an unsupervised manner.

Another well known, more modern classification technique are Support Vector Machines (Vapnik, 1995). These work by maximising the so called margin, the region separating two classes of datapoints in a hyperplane. The datapoints close to the decision boundary (the support vectors) are the hardest to classify and as such play a crucial role in maximising the margin. Other notable classification techniques are Decision Trees (Quinlan, 1986), which classify data through tree-like flow charts in which various features are checked; naïve Bayesian classifiers (e.g.

Friedman et al., 1997), which assign classes to datapoints using Bayes rule while assuming independence between the probabilities of observing different features; and connectionist approaches; the latter being discussed in more detail in section 2.4.4.

2.4.2 Formal Concept Analysis

Formal Concept Analysis (FCA) (Wille, 1982) can be seen as a mathematical formalisation of the classical theory of concepts (Priss, 2006). Originally developed as a sub-field of applied mathematics, FCA proposes a formal treatment of concepts by postulating them as a hierarchically ordered lattice of objects and properties. FCA can be applied to a whole range of fields, including cognitive science, linguistics, data mining and economics, see Ganter et al. (2005) for an overview. Recently, Wennekers (2009) suggested that hierarchies constructed through FCA may have some biological basis. In FCA a triplet $\{O, P, I\}$ (called the context) is considered, where O is a non-empty set of objects, P is an non-empty set of properties and I is a binary relation between O and P indicating whether or not object $o \in O$ exhibits property $p \in P$. A concept is then a pair (A, B), with $A \in O$ and $B \in P$ such that A is the maximal set in O that shares all properties of B and B is the maximal set of properties shared by all the objects in A. The collection of all formal concepts of $\{O, P, I\}$ can be ordered in a hierarchical set-theoretic structure, called the concept lattice, through the requirement that all objects of subordinate concepts are a subset of the set of objects from their superordinate concepts. This yields to a formally sound order which is also intuitively easy to grasp. An example (adapted from Wennekers, 2009) is as follows: $O = \{$tomato, lettuce, spinach, beans$\}$, $P = \{$red, green, veggy, canned$\}$ and I is provided in table 2.2. The resulting concept lattice is displayed in figure 2.1, which depicts the hierarchically ordered set of all (A, B) pairs that can be constructed from context $\{O, P, I\}$.

FCA is mostly in line with the classical theory of concept representation (section 2.3.3): the full set of objects and their properties need to be known in order to build the lattice. This approach clearly suffers from all the objections that were raised against the classical theory, and thus cannot be considered as a proper cog-

	red	green	veggy	canned
tomato	x	0	x	0
lettuce	0	x	x	0
spinach	0	x	x	0
beans	x	0	x	x

Table 2.2: Binary relation between O and P which indicates whether or not object $o \in O$ exhibits property $p \in P$.

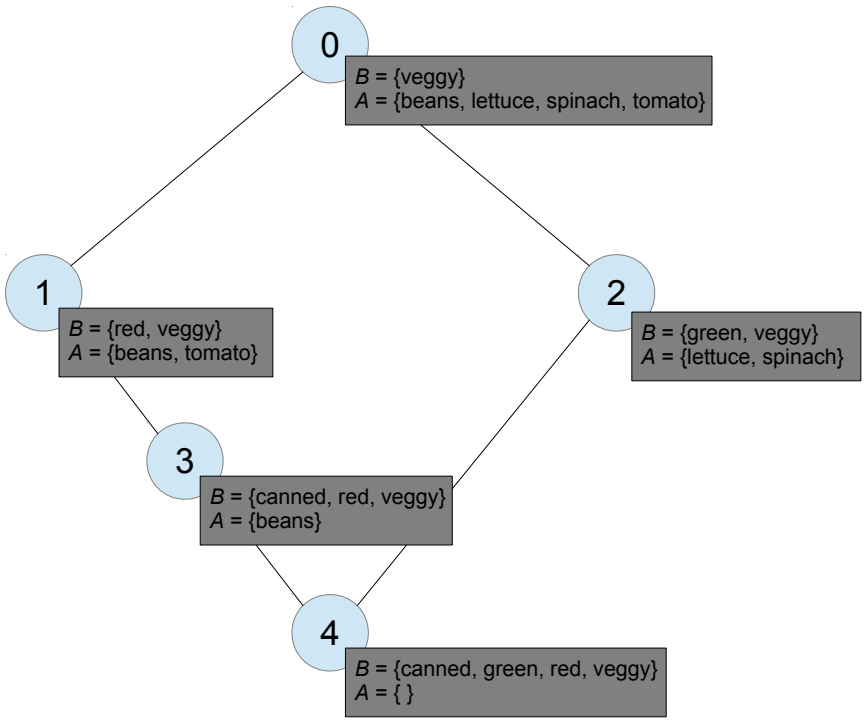


Figure 2.1: Example of a FCA lattice which can be constructed from a context $\{O, P, I\}$.

nitive model (although there have been attempts to reconcile FCA with prototype theory, see Van Eijck and Zwarts, 2004). Nevertheless, the manner of representing concepts as hierarchical ordered sets is intuitively appealing and may help to clarify relations in bodies of data.

2.4.3 Semantic networks and LSA

Another popular method to represent conceptual knowledge is through semantic networks (Sowa, 1991, 1992). This technique is not specific for either psychological or computational accounts of concepts, but could be utilised from both perspectives.

In its core a semantic network is a structure which expresses relations between a collection of objects. Typical relations, at least for definitional networks, are *isa*, *isnot*, *has* and *ispartof*. When properly applied, such relations allow for structured expression of a rich body of conceptual knowledge in a hierarchical fashion. As such, it supports inheritance. Semantic networks can account for typical human memory behaviour like semantic priming; this is discussed in more detail in section 8.3.2.

A problematic aspect of semantic networks is that there is no intrinsic grounding of concepts within the network. The meaning of a concept is purely derived from its relation with other concepts and as such may suffer from the 'dictionary' problem (see section 1.4).

LSA Latent Semantic Analysis as proposed by Landauer and Dumais (1997) is technique for the acquisition of semantic knowledge through statistical analysis of large text corpora. This typicality results in a network that is akin to semantic networks as described above, with the notable difference that the connections between words are not described as logical relations, but rather through associations which carry a certain weight. Throughout their lifetime children are exposed to vast quantities of written and spoken language and as such are able pick up statistical regularities from this. Related words commonly co-occur and by analysing these co-occurrence word meaning can be derived. It has been shown that a system employing LSA can acquire vocabulary knowledge at a rate comparable to school children. While systems trained through LSA may perform comparable on standard word meaning tests, the word knowledge is represented in a multidimensional space with a high number of rather abstract dimensions[8] that cannot easily be related to any kind of modality. For instance, after application of LSA a system might consider 'bread' and 'butter' to be highly related, due to their mutual co-occurrence in conjunction with other related words. This relation is then expressed as a high similarity in the multidimensional space (typically calculated by taking the cosine of the two respective vectors), but it does not say anything about why these two words

[8]The 'magic' number appears to be 300 dimensions, this number is derived through performance tests.

are related; neither the evaluation of the respective vectors provides any insights due to the abstractness of the dimensions. Moreover, a knowledge system derived through LSA lacks any form of grounding, because the meaning of words is purely derived from the relation with other words. In this respect it bears similarities to a semantic network as in both cases the meaning and structure of words is dependent on the network configuration, i.e. the relation to other words. An illustration of the type of networks that can be formed through application of LSA-like techniques is displayed in figure 2.2 (visualised through Gephi[9]). The text corpus that was used for this illustration is the English Lara Corpus from the CHILDES database (MacWhinney, 2000).

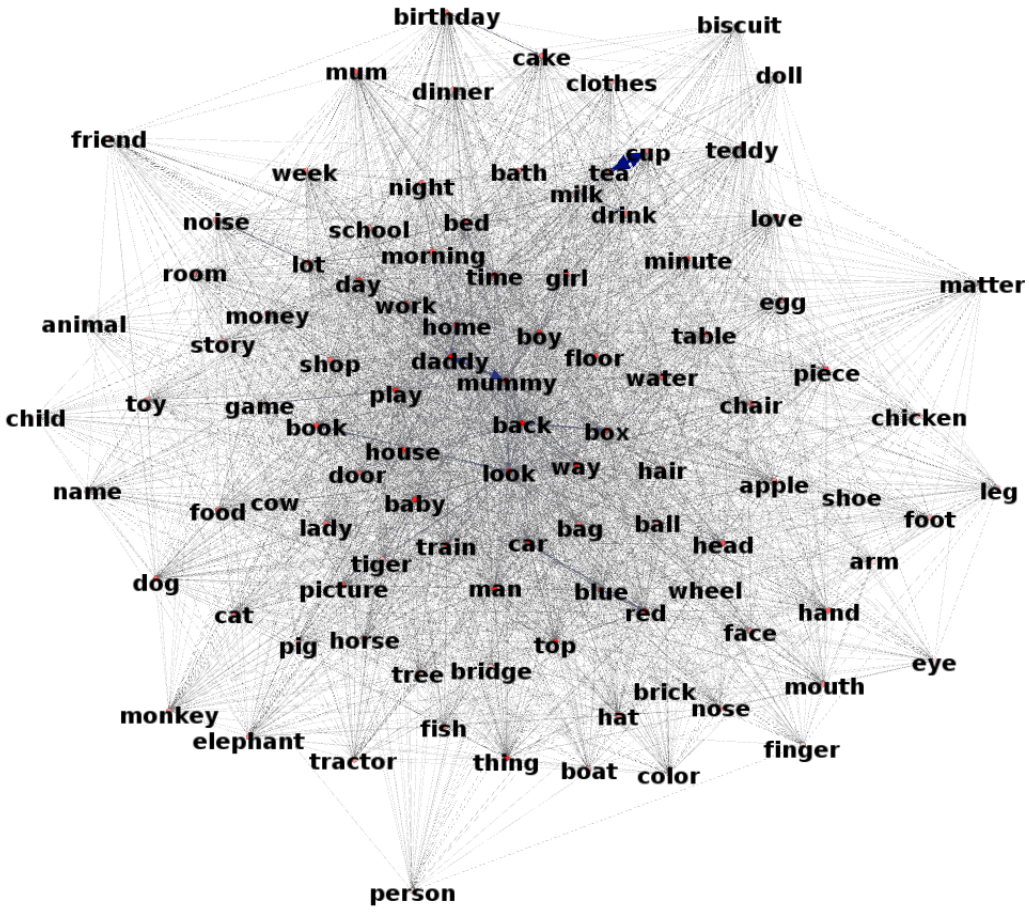


Figure 2.2: Graph displaying thematic relations that were found in the English Lara Corpus from the CHILDES database through application of LSA-like techniques.

[9] https://gephi.org.

2.4.4 Connectionist models

Connectionism is generally thought of as new AI, although some theories and techniques have been around for quite a while and could by now be considered as 'classic new AI'. Nevertheless, as of the time of writing, connectionist models are still being used in a wide variety of applications. In relation to conceptual modelling and the theories from the human perspective described previously, connectionist models tend to produce classification behaviour that is mostly in line with prototype and exemplar models, e.g. (sets of) output units representing different categories, with their activation levels providing a measure of similarity between input data and the respective categories.

Inspired by neurological accounts of cognition, connectionism is a family of theories that advocate the notion of distributed representations rather then explicit ones. Connectionist approaches typically provide mappings between sensory input and classifying output; that is, the output nodes can be read as classification of whatever input the network has received, and this in turn can trigger various behaviours. The biological analogy is as follows: in natural cognitive systems the senses provide input and the behaviour of the organism is the output. In between the brain engages in some kind of cognitive computation. The brain is a massively parallel structure and on a low neurological level, i.e. individual neurons, there appears to be no explicit representation of chunks of information. Rather, the information exists as the collection of hundreds/thousands/millions of neurons which collectively store information in a distributed fashion. In a similar vein artificial neural networks process information in parallel. Sensory input results in specific output patterns as a reflection of the connectively of the neural network, i.e. how neurons are interconnected and what the strength (weight) of these connections is. Some common neural networks and their ability to classify input spaces are discussed below.

Even the most simple form of a feed-forward neural network (FFNN), the perceptron, is able to function as a classifying mechanism because, after sufficient (supervised) training, it is able to classify an input vector as belonging to a certain class. For a perceptron to be able to learn to distinguish different classes, they need

to be linearly separable[10] (Rojas, 1996). Recurrent neural networks (RNN) (Elman, 1990; Jordan, 1997) are a more advanced form of neural networks due to the addition of recurrent connections which allow for temporal storage of data and reuse in subsequent time steps.

A self organising map (SOM) (Kohonen, 1982, 1984) is a type of neural network that originally was proposed as a method of visualising high dimensional data into a (typical) 2D grid. Its functioning is akin to FFNNs, but it has the additional feature of topology preservation. This means that the output nodes are organised in a coherent lattice, such that output nodes that respond to similar data are neighbouring each other. Hence, input vectors that are sufficiently similar will result in neighbouring output nodes being activated. SOMs can be used as classifiers; they can reduce high dimensional input data into meaningful classes on a 2D grid.

Interactive Activation and Competition (IAC) models (McClelland and Rumelhart, 1981; Rumelhart and McClelland, 1982) constitute another flavour of neural networks with specific classification properties. IAC models consist of a network of multiple pools of neurons, with each node having a certain activation value and activation is propagated through the network via weighted connections. Different pools typically encode different modalities, and within these pools individual nodes encode specific properties, thus constituting localised representations (Page, 2000). However, rather than having each node potentially connecting to every other node, within IAC models all the nodes in one pool have inhibitory connections to one another, while they have excitatory connections to special 'hub' nodes. As such, these hub nodes serve as a bridge which connect multiple modalities. Because the connections are bidirectional, the activation of nodes in a certain pool can spread via a hub to other pools. In this respect IAC models bear similarities to concept-like behaviour, as the perception of certain properties can trigger associations with related properties.

IAC models can account for various properties of human memory, for example priming effects, spreading activation and top-down effects. With respect to con-

[10]Data is said to be linear separable when for each n-dimensional space there exist a hyperplane of n-1 dimensions that cuts this space in two, as such separating data points into distinct groups.

cept modelling, the property of localised representations allows for the formation of knowledge structures that are interpretable in a concept-like manner. Particularly the hub nodes can be interpreted as prototypes, as the connectivity of the hubs and the activation levels of connected nodes in different pools specify which properties a prototype tend to have. Hence, the observation of an object that is very similar to a known prototype results in high activation levels and clearly distinctive properties, while observations that are less similar to known prototypes result in less articulated activation patterns. A major drawback has been that traditional IAC models did not incorporate any form of learning but had to be designed by hand. Recognising this as a considerable drawback, augmentations which allow learning in IAC models were proposed, for instance by Burton (1994).

A relative new proposal for a cognitive model, Epigenetic Robotics Architecture (ERA) (Morse et al., 2010) combines several aspects of connectionist models described above. The basic units of representation are based on self-organising maps; that is, different SOMs allow for the encoding of information in different modalities. Multiple SOMs are connected to one hub, which typically encodes body postures. These hub SOMs that combine different modalities through other SOMs constitute the basic ERA units; as such this is very much akin to the hub structure of IAC models. Multiple ERA units can be combined in hierarchical fashion, allowing for more abstract representations. Linguistic labels are associated with activation patterns in the ERA unit through Hebbian learning (Hebb, 1949).

2.4.4.1 Modelling concepts

What all connectionist models tend to have in common, is that they provide a mapping from certain input patterns to specific output patterns. Thus, they are suitable as classifiers. This mapping is typically the product of a training procedure, either in supervised (e.g. FFNN, RNN) or unsupervised (e.g. SOM) fashion.

Various studies have shown that connectionist models can be used to learn and represent categories in such a way that they can account for characteristics and empirical findings associated with prototype and exemplar models. For instance, Reilly

et al. (1982) trained neural networks to classify unconstrained handwritten numerals through the formation of prototypes in the system's memory. The implementation was a middle ground between prototype and exemplar theory, as every exemplar is treated as a prototype. Also Schyns (1991) proposed a model that is based on an underlying connectionist approach (a SOM in this case) which is able to learn categories. This model made a distinction between the learning of perceptual categories on the one hand and the associated linguistic labels on the other. There have also been attempts to combine connectionist models with the theory-theory of concepts, for instance in (Towell and Shavlik, 1994) neural network learning mechanisms were augmented with "domain theories" represented in propositional logic.

A classical drawback of some forms of connectionist modelling is the fact that they suffer from catastrophic interference; that is, the problem of forgetting what was already known when new data is learned (see French, 1999). Another issue is the fact that knowledge is represented in a distributed fashion. This makes it rather hard to analyse and identify the internal workings of these networks and to gain an understanding of the basis on which classifications happen. In contrast, IAC networks (described above) do not exhibit this problem because of their localised manner of representation.

2.4.5 Concepts and word labels

Regarding the relation between concepts and words as discussed in section 2.3.1, the different approaches to modelling concepts from the machine perspective discussed above treat this relation quite differently. First of all, in FCA with its definitional structure concepts are logical constructs that are defined by their properties. As such, there is no distinction between a concept and a word, as a concept is in fact identified by a word. The associated properties are all members of logical concept set. Secondly, network-like structures such as semantic networks and LSA could be said to be devoid of concept-like entities (other than words), as they are purely constructed by word-word relations. Words *are* the concepts, there is no other level of representation, and the meaning of words is derived from the relation to other

words. Thirdly, connectionist models typically do treat word labels as independent from conceptual levels, as the latter are usually realised through sensory activation patterns. As such, the network forms associations between word labels and sensory input. Word labels are the output of the classification process, and with respect to learning are the basis on which different sensory patterns can be clustered. In case of unsupervised forms of learning (e.g. SOMs), the result is a structuring of the input space which typically lacks word labels. In this case subsequent association between conceptual representations and word labels typically does involve some other learning mechanism that might be supervised or more implicit, e.g. Hebbian learning.

With respect to modelling concepts, it is not the distributed nature of connectionism that is interesting (indeed, this may be problematic to a certain extend, e.g. categorical forgetting), but the associative nature between sensory patterns and word labels that seems most suitable for concept modelling. The construction of these associations might best be achieved by subscribing to a theory of social and interactive learning.

2.5 Modelling concept learning

In the sections above some theories concerning the use of concepts have been discussed. From a human perspective these studies of concept use tend to be heavily intertwined with theories about learning and memory. Commonly, the different explanations are concerned with the particular structure that human concepts have; the fact that people use concepts is treated as a given, the theories seek to explain what kind of underlying structure gives rise to this phenomenon and what kind of psychological, neurological and/or sociological factors influence this. The machine perspective on the other hand is mostly concerned with the question of how (abstract) knowledge can be represented and how an automated system can learn this kind of knowledge. Human concept use is viewed as a very effective way of dealing with a vast amount of information coming from a constantly changing, dynamic environment. As such, from the machine perspective, the focus lies on finding for-

malisms and algorithms that allow for effective ways of representing and learning knowledge. These do not necessary have to be compatible with human concept use, but the fact that people prove to be exceptionally good at this provides at least one solution from which inspiration can be drawn.

For both the human and the machine perspectives the notion of social learning (section 1.5.1) is deemed important, as the modelling of concepts and specifically the dynamics of acquisition needs to be viewed in a social context. That is to say, from the human perspective a theory disregarding social aspects would be incomplete and from the machine perspective there is a great potential with respect to learning based on social interaction. As such, a perspective on concept learning that takes these social aspects into account seems a promising endeavour.

Combining insights from the human perspective and the machine perspective, and endowing these with the notion of social learing, the aim is to arrive at a functional model of social concept learning that can be embedded on a robot so that it can learn through interaction with people. What then, in order to build such a computational model of concept learning for social HRI, are the fundamental aspects that need to be taken into account? Broadly speaking, an effective computational model will need the following:

1. A representation of conceptual knowledge in such a way that the properties of this model, including acquisition mechanisms, sufficiently reflects empirical data from human studies. That is, the representation of concepts should not be in contradiction with anything that is known from psychological studies of human concepts, but should be able to account for properties of human concept use (as discussed in section 2.3).

2. A means of interacting with its social environment that is perceived as natural by people, as to support the learning of concepts through social interaction between a robot and a person.

Addressing 1), following a relatively modern approach towards conceptual modelling from the human perspective, it seems prudent to account for the prototypically

of concepts (section 2.3.5). Indeed, as Murphy concludes in his final chapter, "I will propose that our theory of concepts must be primarily prototype-based." (Murphy, 2002, p.488). This conclusion is based on the fact that prototype theories seem to be able to account for most of the empirical data (better than the exemplar approach); albeit while being part of a larger knowledge structure as envisioned by the theory-theory (section 2.3.6). While prototype theory is by no means the 'ultimate' theory of human concept use, its properties and explanatory power are nevertheless considered adequate for the type of research that is pursued in here; that is, interactive learning by embodied agents. A relatively modern approach that has its basis in prototype theory is the framework of conceptual spaces.

Addressing 2), acknowledging that learning always occurs in a social context through interaction between learner and a teacher, and the observation that concept learning is heavily intertwined and indeed, is quite frequently indistinguishable from, the learning of words (section 2.3.1), a minimal model of word learning through social interaction is deemed appropriate. Such a model is realised through the adoption of the language games framework.

The combination of concept representation based on conceptual spaces and the language game framework is not novel in itself, as various studies have used a similar setup. These will be discussed in more detail in section 3.3. The aim of this work however, is to use these frameworks to investigate some interesting topics regarding social concept learning (as outlined in section 1.6), using both simulations and an HRI setup. The conceptual spaces and language games frameworks will be described in more detail in chapter 3, along with a description of related literature using these frameworks and implementation details as to obtain a computational model which serves as basis for most experiments reported in this thesis.

2.6 Chapter summary

This chapter has discussed concepts, both as part of human cognition and how they might be modelled from a perspective of artificial intelligence. From the perspective of human cognition most theories of concepts are concerned with the explanation of

empirical data from psychological experiments by providing models of conceptual structures that may give rise to this data. The machine perspective on the other hand is less concerned with explanation of empirical data but places more emphasis on the construction of models that might enable artificial systems to use concepts in a human-like manner.

Various psychological theories of human concept use have been discussed and they were linked to some approaches from the machine perspective. A synthesis was provided of how a computational approach might yield a model that is sufficiently compatible with relevant psychological theories. It was concluded that for the aims of this work, that is, the social learning of concepts for embodied artificial agents, an approach based on prototype theory might be most appropriate in terms of capturing sufficient aspects of human concept use while at the same time providing an account that is relatively straightforward to incorporate in computational models of agents and an HRI setup. In this approach concepts are modelled as entities that are separate from the words that describe them, while at the same time linguistic interaction governs the manner in which concepts are learned. The need to coordinate such social linguistic interactions that result in concept learning by agents was acknowledged by adopting an appropriate framework to model this.

Chapter 3

Computational model

In this chapter we describe the computational model which forms the basis of most experiments reported in this thesis. First a description is provided of two theoretical foundations on which the model is based, namely conceptual spaces and language games. We then illustrate how these two theories can be combined into a functional model. Simply put, a computational model of concepts is envisioned as consisting of two parts: 1) a low level classifier that can process incoming stimuli into discrete chunks, based on their resemblance to known patterns; 2) a higher level organisational structure that links different chunks from the lower level together, based on co-occurrence and governed through linguistic interaction (the latter bears similarity to Hebbian learning). More specifically, a conceptual space serves as a way to represent concepts within different domains, and the language games provide a manner for acquisition of linguistic descriptions for these concepts. Thus, there exists a difference between the (perceptual) concept on the one hand, and the linguistic description (word) an agent may use for this particular concept on the other. For a discussion about the relation between words and concepts, see section 2.3.1.

3.1 Conceptual Spaces

A long standing debate within cognitive science is whether knowledge representation can best be understood in terms of high level symbolic representations or in terms of associative distributed structures (usually dubbed connectionism). See section 1.2.1

for a more in-depth account. Rather than championing one view or the other, the conceptual spaces (CS) framework (Gärdenfors, 2000a) proposes to represent knowledge on a level that resides *between* symbolic representation and distributed representation. In this view, connectionist models are seen as subconceptual, detailed processing of the lowest level of information units, while symbolic processing is seen as the most abstract form, providing higher level computation and logical reasoning. Thus, symbolic representations and distributed representations are not seen as competing theories, but rather as explanations of the the same phenomenon on different levels. As such, a CS is placed in between these levels, describing concepts in terms of geometrical shapes that are grounded in sensory properties which can be distributed in nature, but which can also exhibit symbol-like behaviours in more abstract levels of description.

A CS consists of a geometrical structure in vector space that represents a collection of one of more domains, e.g. colour, size, taste. Each domain consists of one or more inseparable quality dimensions that represent the lowest level values of the particular domain. For instance, the colour domain is typically represented using three dimensions, RGB being a popular choice[1]. To express a colour, each of the three quality dimensions needs to have a value; expression of colour using only R and G would be impossible. As such, inseparable dimensions together form a domain. In similar fashion, the taste domain can be represented through four dimensions encoding values for sweet, sour, bitter and saline (or five dimensions if umami is included).

In principle, any domain can be used to express a concept in a CS, provided the relevant dimensions can be found. For some domains it is more straightforward to extract the relevant dimensions than for others. For instance, in the domains discussed so far associated dimensions have perceivable and measurable values, but the shape domain might include a dimension like 'cat-like' for which it is harder to express explicit values. However, often dimensions will relate to perception; the

[1]RGB is commonly used to represent colours on display devices, but there exist a wide range of colour models with different properties and virtues. E.g. HSV, CMYK, CIE-XYZ and CIE-L*a*b*, the latter being close to human colour perception. The majority of these spaces appear to consist of 3 dimensions.

human senses provide domains (vision, smell, touch, hearing, taste) and their origins are innate (Gärdenfors, 2000a, p. 27). In other cases the dimensions may be more analytic, i.e. the weight of an object (which cannot directly be perceived).

Crucial to modelling concepts in a CS is the ability to take a distance measurement. For each of the dimensions involved, a suitable metric to calculate distance must be defined. For a lot of dimensions the Euclidean distance may be the most appropriate one, but the Manhattan distance can also be used[2]. The metric can be augmented with a weight to allow certain dimensions to be more prominently expressed than others; see equation (3.1). Different contexts can influence saliency and this saliency determines the dimension's weight.

Distance d_{xy} between point x and point y takes the general form:

$$d_{xy} = \left(\sum_{i=1}^{N} w_i |x_i - y_i|^r \right)^{\frac{1}{r}} \tag{3.1}$$

where r denotes the type of metric with $r = 1$ for the Manhattan distance and $r = 2$ for the Euclidean distance, i is a dimension and w the weight of the dimension.

This distance measurement allows for similarity judgement, as similarity between two points in the CS are a function of their distance. However, with respect to similarity judgement, studies have shown that people do not rate concept similarity in a linear fashion, but rather there appears to be a non-linear relation between distance and similarity (Shepard, 1987; Nosofsky, 1986). To account for this, distance is converted into a similarity as expressed in equation (3.2). Similarity s between i and j is computed as an exponentially decaying function of distance:

$$s_{ij} = e^{-cd_{ij}} \tag{3.2}$$

where c is a sensitivity parameter.

Rather than mere points in space, concepts are postulated as convex shapes which may span one or more domains. In each dimension of the domains that are associated with a particular concept, the coordinates of the concept in this dimension

[2]Typically the Euclidean distance is used for distance measurement on continuous dimensions and the Manhattan distance is used for discrete dimensions.

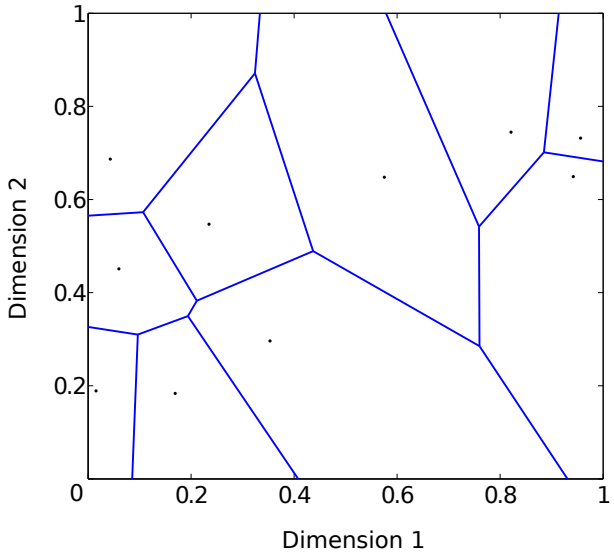


Figure 3.1: Illustration of a simple conceptual space with 2 dimensions which is populated by 10 concepts. Through generation of a Voronoi diagram the boundaries of the concepts are defined.

are used to generate a Voronoi tessellation. This results in a decomposition of the space as depicted in figure 3.1 which shows a simple CS with two dimensions that is populated by 10 concepts. Each point in the space is assigned as belonging to the concept to which it is closest and as such the conceptual boundaries are defined. The region encompassed within the generated polygon surrounding a concept is said to be part of that particular concept.

3.1.1 Populating the CS

Typically a learning agent starts out with an empty CS, i.e. without any conceptual knowledge. Through interaction with other agents concepts can be learned in, from the learning agent's point of view, essentially a supervised manner[3]; a newly observed stimulus can either be stored as a new concept in the CS, or added to an existing one. The decision of whether to store the new data as a new concept, or blend it in with existing concepts is based on language game dynamics which take place outside of the CS. This is described in detail in section 3.2.

Once a CS is populated with some concepts, it can be used to classify new

[3]From a language games perspective the learning is not supervised but rather exhibits self-organising properties, as a population of agents generate shared meaning through interaction with each other, see section 3.2.4.

observations. Classification happens in the following manner. Given a CS C with concepts $< c_1, c_2, ..., c_n >$, each $c \in C$ consists of a collection of one or more domains $D^c =< d_1^c, d_2^c, ..., d_n^c >$ and each $d^c \in D^c$ consists of one or more quality dimensions. For a newly observed stimulus y with associated domains D^y, similarity s_{cy} is calculated for each $c \in C$. y is then classified as c for which s is the smallest. Based on certain thresholds y can either be stored as a new concept in which case y is stored as c_{new} in C, or y can be added to the closest matching $c \in C$. In case of the latter this is done as follows: for each $d^c \in D^c$ that is equal to $d^y \in D^y$, the coordinates x of d^c are updated through inclusion of the coordinates z of d^y. This is done using the following formula:

$$x_{new} = x_{old} + \frac{|x_{old} - z|}{N} \tag{3.3}$$

where N is the number of observations on which x is based, so after the update operation $N = N + 1$. In a similar fashion, the standard deviation for each $x \in d^c$ is updated as well. The SD of x provides a measure of spread for this dimension, and as such is an indication of the relative size of the concept in the CS.

3.1.2 Prototypes and exemplars

As described in chapter 2, a prototype based manner of concept modelling seems to be sufficiently close to empirical data collected from human studies. A CS representation is very well suited for prototype modelling, as the inherent distance metric can easily function as a measure of typicality. The model's ability to represent prototypes is further explored in section 4.4, which provides a practical example of prototype learning featuring a commonly used dataset of zoo animals.

Even though a CS is typically populated with summary representations as concepts and as such supports a prototype approach, it can easily be extended to accommodate aspects from the exemplar theory (section 2.3.5), by simply keeping a copy of all observed stimuli in memory. Indeed, this was the case for the actual implementation that was used for experiments reported in this thesis. Although not explicitly used, the collection of all observed stimuli that make up the prototype is useful for calculating the SD of the prototype, thus gaining a measure of the spread

of a concept within the CS. Should the need arise (e.g. because of some vital empirical data strongly suggesting exemplar based modelling), the distance measurement can be modified through utilising the set of exemplars rather than the prototype. As such, the CS model implemented for the work reported in this thesis also supports concept modelling based on exemplar theory.

3.1.3 Related work

The CS framework has been applied in numerous other works, varying from theoretical expansions to more practically oriented applications in robotics. An extension, which describes how concepts can also be represented in an 'action space' is described in Gärdenfors (2007). More practical examples of robotic applications which are geared towards categorical perception, are provide by Chella et al. (1997, 2000). In these works a robot vision system is based on conceptual spaces in which each point corresponds to a geon-like 3D geometric primitive (Biederman, 1985).

A notable criticism of the conceptual spaces framework is that application of the theory has only been demonstrated in simplistic cases (Tanasescu, 2007). In response to this, Adams and Raubal (2009) introduced a more comprehensive conceptual space algebra which provides query operations for semantic similarity measurement and concept combination, as to "allow one to build and reason with complex conceptual space structures". Further extensions of conceptual spaces, including a symbolic subspace which allows for a connection with higher order symbolic representations, are described by Aisbett and Gibbon (2001) and Rickard et al. (2007).

3.2 Language Games

The second part of the model is based on language games; this serves as a mechanism to govern the acquisition and association of linguistic labels with perceptual concepts that populate the CS. Because language games are inherently placed within a social context, that is, agents learn through interaction with other agents, this framework supports a social learning account as discussed in section 1.5.1.

3.2.1 Background

The language games model was first proposed by Steels (1996a,b) as a model of meaning acquisition for artificial agents. A real world analogy, adapted from (Steels, 1999) is described below.

Imagine the following scene. Two people with different native languages, say French and English, are dining together. At some point the English person asks the French "Could you pass me the salt please?". The French person might not be very accomplished in the English language, but may nevertheless register the fact that the statement is a question. Furthermore, "salt" sounds somewhat similar to the French "sel" and given the situation it is not too hard to deduce the meaning of the question. Thus, the French person passes the salt and the communication is a success, which can be expressed in a non-verbal manner. Both interlocutors mentally note the success of this interaction and the French person can also mentally strengthen the connection between the word "salt" and the salt shaker he just passed. Alternatively, it might be the case that the question is not understood properly. The French person might indicate this through some kind of non-verbal action, or simply fail to pass the salt. If this happens the English person might make a gesture towards the salt on the table and utter the word "salt" again. Even though the initial communication was not successful, the resulting actions still allow the French person to mentally connect the salt on the table with the linguistic label "salt".

Through a continuous engagement in these kind of ritualised interactions, interlocutors can gradually learn to associate the correct linguistic labels to objects in their environment, thus effectively learning to speak a shared language. It is these kind of interactions that Steels argues help to establish meaning between abstract word utterances, real world objects and situations. In a situated environment agents exchange linguistic descriptions of objects they perceive through communicative acts. Meta-communicative actions (which can be verbal or non-verbal), expressing confirmation or disconfirmation of these communicative acts, provide the agents with means to alter the associations between the linguistic expressions and their referent, i.e. the objects in the world. Various software simulations using this

model have shown that through prolonged engagement in language game interactions, communication between agents tends to become more successful over time, i.e. they converge to a shared system of meanings (Steels and Kaplan, 2002b; Steels, 2003).

Language games may be seen as an umbrella term, as there exist a multitude of variations and implementations. The underlying principle is that agents develop a shared language that enables them to 'speak' about their environment, through the exchange of linguistic labels and the adoption and shifting of underlying conceptual representations; variations differ in the number of agents, the manner of interacting, the environment, the type of feedback between agents, the agent's update functions etc.

Experiments in language games can be done in simulation, but have also been extended to robotic hardware. For instance, the Talking Heads experiment (Steels, 1999) included pan-tilt cameras that observed a real world scene of colours and shapes. Other experiments included various robotic platforms such as Lego robots (Steels and Vogt, 1997), AIBOs (Steels and Kaplan, 2002a), QRIOs (Wellens et al., 2008; Steels and Spranger, 2008) or human subjects (Belpaeme, 2002a). The model has also been used to explore the emergence of human communication (Steels, 2006), compositional structures (Vogt, 2005) and linguistic categories (Puglisi et al., 2008), as well as case studies focussing on the origin of colour terms (Steels and Belpaeme, 2005; Belpaeme and Bleys, 2005).

The implementation that is used in this thesis follows the model as described in Belpaeme and Bleys (2005). As such, we take the following view. Language games consist of two distinctive parts: the *discrimination game* and the *guessing game*. Discrimination games are used by the agent to develop a categorical representation from the input space; guessing games serve to develop linguistic descriptions for these categories through interaction with other agents in a population. As such, after engaging in language games agents are able to successfully communicate with each other about their shared environment. The agents, the discrimination game and the guessing game are described in more detail in the following sections.

3.2.2 The agents

Language game always involve one or more agents. An agent A is defined as follows:

$$A = \{S, L, M_{i,j}\}$$

where S is a conceptual space that is populated with i concepts, L is a lexicon containing j word labels and $M_{i,j}$ is a $i \times j$ matrix encoding connection strength between all $c \in S$ and all $l \in L$ as a scalar $[0.0, 1.0]$. When multiple agents (N) are involved, they are collected in a population $P = \{A_1, A_2, ...A_N\}$.

3.2.3 The discrimination game

The discrimination game is played by an individual agent observing a context, which is a set of objects from the agent's environment. It proceeds as follows:

1. Agent A observes context $O = \{o_1, ..., o_N\}$ containing N objects and an index i, designating object $o \in O$ as o_t: the topic of the discrimination game.

2. A finds the best matching concept c from its conceptual space S_A for each stimulus in the context: $\{o_1, ..., o_N\} \rightarrow C = \{c_1, ..., c_N\}$.

3. For o_t, if the best matching concept c_t is unique in C the game succeeds, otherwise it fails.

The discrimination game can fail in several ways; this is an opportunity to improve the agent's concepts. When S_A is empty, a new category is created on the coordinates of o_t and the game proceeds with the next round. When no unique discriminating concepts can be found for o_t, there are two possible actions: (1) a new concept is created on o_t, or (2) the best matching concept c is adapted to better represent o_t by shifting c towards o_t. Action (1) is taken when the discriminative success[4] of the agent is below a threshold $adapt = 0.9$, otherwise action (2) is taken. Regardless

[4]The discriminative success of an agent is the global success of the agent in all discrimination games it has engaged in. It is typically measured by dividing the number of times the agent has successfully discriminated the topic from the context by the total number of discrimination games the agent has played.

of which action is taken, the discrimination game results in A assigning a concept from S_A (either new or adapted) as best matching o_t.

A division of input space into discrete categories as achieved by the discrimination game can be compared to other methods of categorisation, like k-Means and self-organising maps. This comparison is covered in more depth in section 4.3.

3.2.4 The guessing game

The guessing game is played between two agents observing the same context. One agent (the teacher) initiates the game by describing an object from a set of multiple objects (the context), and the other agent (the learner) tries to guess which object from the context the teacher has described[5]. More formally, the guessing game proceeds as follows:

1. Teacher A^T and learner A^L observe context $O = \{o_1, ..., o_N\}$ containing N objects and the index of the topic, specifying o_t.
2. A^T plays a discrimination game for o_t; this results in the concept c^T.
3. A^T finds the associated label l^T and communicates this to A^L.
4. A^L hears $l^T = l^L$ and finds the associated concept c^L.
5. A^L points to o^L closest to c^L.
6. if $o^L = o_t$, the guessing game succeeds; if not, it fails.

When the guessing game is successful, the connection strength between l^L and c^L is increased by learning-rate α and o_t is added as an exemplar of c^L, effectively shifting the coordinates of c^L towards o_t.

The guessing game can fail in several ways. (1) The discrimination game of A^T fails; in this case the guessing game fails as well. (2) A^L does not know l^T. A^L then plays a discrimination game for o_t, finds c^L and adds $l^T \rightarrow l^L$ to its lexicon

[5]Throughout this research the agents are called teacher (the agent that initiates the guessing game) and learner (the agent that makes the guess), as we are mostly concerned with a teacher-learner scenario. In related research that utilises the language games framework, agents are also called speaker and hearer. Despite this difference in terminology, the respective roles of the agents are the same.

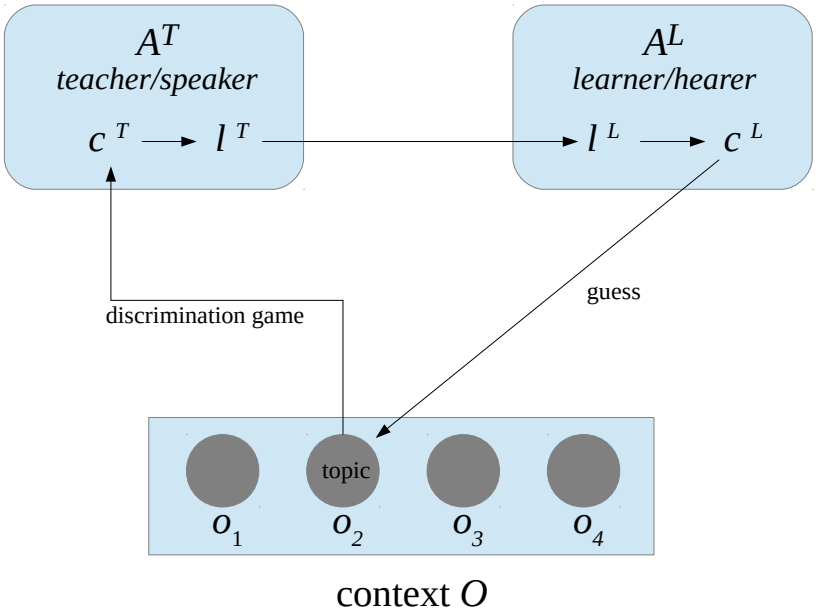


Figure 3.2: Schematic display of the guessing game interaction between a teaching agent A^T and a learning agent A^L according to the description provided in section 3.2.4.

with a default connection 0.5 to c^L. (3) A^L knows l^T, but points to the wrong topic. A^L then decreases the connection strength by α between l^L and c^L, plays a discrimination game for o_t, finds c^L and adds $l^T \rightarrow l^L$ to its lexicon with a default connection 0.5 to c^L. A schematic representation of this interaction is displayed in figure 3.2.

3.2.5 Two scenarios

There are two different scenarios possible with respect to playing language games that involve multiple agents. The first one resembles a teacher-learner scenario, in which one agent (the teacher) is initialised with certain categories and words combinations, allowing it to describe objects from a context. As such, the training data which is used during guessing games will be of a similar structure (same domains and dimensions) as the knowledge the teaching agent is endowed with. The other agent starts with a blank memory, i.e. its lexicon and conceptual space are empty. Both teacher and learner engage in a series of guessing games, and through these interactions the learner gradually acquires knowledge from the teacher. The teacher

does not modify its word-meaning associations.

The other variety of language games is played in a population of two or more agents. In this scenario all agents start with a blank memory. During each round, two agents that are randomly selected from the population play a guessing game. If words are needed to describe the topic from the context, the agents acquire these through a discrimination game. Through multiple interactions between all agents, word-meaning combinations gradually propagate throughout the population, eventually converging to a shared language. This latter type of language game is mostly applied when one is interested in the dynamics of language evolution.

3.2.6 Evaluation

To measure the performance of the agents in terms of learning, i.e. how effective knowledge from a teaching agent can be 'transferred' onto the learning agent, or in the case of a population, how well do the agents in the population 'speak' the same language, a performance measurement needs to be defined. Typically this is done by measuring the communicative success of language games. The communicative success is calculated as the number of language games in which two interacting agents are successful in their communication divided between the total number of language games played. Related work featuring the language game framework sometimes measures only the average success in communication over the last x number of interactions, i.e. a running average. However, in this work communicative success is always averaged over all interactions. As such, agents can never reach 100% success, because inevitably there will be miscommunication, particularly at the early stages of interaction when agents' lexicons are not yet aligned to a great extend.

The choice of using this overall measure of communicative success as opposed to a running average was arbitrary, i.e. the practice was simply adopted while implementing the language game framework. It is not expected to influence the experimental findings in any way.

3.2.7 Parameters

Various parameters govern the outcome of language game interaction. Depending on the objective of the experiments, optimal settings will vary. A description of the most important parameters is provided below:

- **Number-of-agents** (N_A)

  This determines the number of agents that participate in the language game. As described above, a teacher-learner scenario is normally played with two agents; other types of language games have two or more agents in the population.

- **Number-of-interactions-per-agent** (N_I)

  This specifies the average number of language game interactions an agent participates in. For convergence of meaning within a population N_I is typicality set to 10,000 games or more, depending on the complexity of the environment, as expressed by other parameters. However, particularly for teacher-learner scenarios, the number of interactions might be considerably less, e.g. some of the experiments that are described in section 7.4 only use 50 interactions.

- **Context-size** (C)

  The context size determines how many objects are included in a context. The higher this number, the harder it is for other agents to make the right guess when observing a context and hearing a linguistic utterance from the speaking agent. Values might range between [2, 7]; for most experiments reported in this thesis C is set to 3 or 4.

- **Minimum-distance-objects** (D_{min})

  Related to C is a measure of how different the objects in the context are with respect to each other. Even with $C = 2$, if these objects are very similar, it is hard for agents to make a guess that is above chance. D_{min} governs the extent to which objects in the context are sufficiently different, so that agents can discern between them. In a human context, this parameter relates to the notion of 'Just Noticeable Difference', i.e. the degree at which two stimuli differ

in such a way that people can just notice this difference. See Baronchelli et al. (2010) for an exploration of this notion.

- **Adaptive-threshold** (T_{adapt})

  This parameter determines how adaptive agents are with respect to building their categories during discrimination games. T_{adapt} functions as a threshold; if the discriminating success of an agent drops below it, the agent invents a new category. A high value (i.e. 95%) will cause an agent to build many categories, while a low value will cause an agent to end up with few, but potentially more general categories. The most effective value for T_{adapt} varies for different language game scenarios.

- **Learning-rate** (α)

  The learning rate specifies the amount of adjustment agents make in their word-category scores, based on the outcome of a guessing game. A high learning rate can result in quick convergence of a certain word-category combination, but this can be disruptive on a population level as it might cause large fluctuations in the shared lexicon.

- **Number-of-replications** (R)

  R encodes the number of replications of the same simulation. Because of random aspects in the initialisation of training data, the selection of agents for guessing games and the choice of topic from the context, the outcome of language game simulations will vary to a certain extent from run to run. To obtain an average measure for statistical analysis, simulations with the same parameter settings are replicated a number of times. Through a trial and error process it was established that $R = 25$ generally provides a middle ground between obtaining good average values and the computational time required to do so.

3.3 Combining CS and LG

The combination of a conceptual spaces model of concept representations (section 3.1) with the language games dynamics that model the development of a shared system of meanings between agents (section 3.2) will form the basis of the computational models that will be described in the following chapters. There exist a body of work in the literature that has followed a comparable approach to concept modelling, a description of this related work is provided here.

Originally, categories in language games were represented through discrimination trees (Steels, 1996a). Following work utilised other means of representation as well, such adaptive subspaces (De Jong, 1999), radial basis function networks (Steels and Belpaeme, 2005) or vector spaces that, although not always explicitly mentioned, bear considerable resemblance to Gärdenfors' conceptual spaces model; e.g. Belpaeme and Bleys (2005); Spranger et al. (2010); De Beule and Bleys (2010), as well as various experiments reported in Steels (2012). Also in Vogt (2005), a conceptual spaces model is used to represent concepts (colours and shapes). In this work the emergence of compositional structures are explored by combining the underlying dimension of the CS with a rewrite grammar that allows for compositional structures. Similar work is described by Vogt (2004), in which agents acquire categories based on principles of cross-situational learning (Siskind, 1996; Smith, 2005; Smith et al., 2006); see section 8.2 for some more discussion of topic. Other approaches employ similar manners of category representation and transfer of meaning, but utilise affordances for categorisation and cross-situational learning, e.g Takáč (2008). Yet other studies have returned to discrimination trees as a means of representing concepts (Wellens et al., 2008).

As illustrated by the related literature described above, the combination of Language Games with an conceptual spaces manner of representing concepts is not novel in itself. However, in this thesis we will utilise this combination as a means to address the questions as formulated in section 1.6, approaching the topics of concept learning and linguistic convergence from the perspective of concept learning through social interaction. As such, the following aspects of this work can be considered as

novel: an explicit relation of the model's concept formation dynamics to prototype theory (section 4.4), the investigation of interactive features as a means to improve learning dynamics (chapter 5), an illustration of how application of the model's dynamics might help overcome differences in embodiment (chapter 6), and the subsequent implementation of aforementioned interactive features into augmentations of an embodied agent's social repertoire (section 7.4).

3.4 Chapter summary

In chapter 3 a description has been provided of the two main frameworks that serve as a basis for the experimental work covered in this thesis. The first framework is that of conceptual spaces, which provides the means for modelling concepts in a prototype-like fashion. Concepts are depicted as convex regions in a multidimensional feature space, and as such the notion of similarity between different concepts (points in the space) is realised through a distance metric. Furthermore, it has been described how a conceptual space can be populated with conceptual structures, and how these structures exhibit properties that are compatible with a prototype theory of concepts.

The second foundation of the experiments reported in this thesis is the language game framework. This functions as a learning mechanism through which a conceptual space of an agent can become populated with concepts and word labels to describe these concepts can be obtained. The language game framework emphases the interactive nature of learning, as the meaning of objects in an environment gradually becomes aligned within a population of agents through multiple interactions. As such, these learning mechanisms provide a ground for social interactive learning, both in simulated environments with multiple agents, and in a didactic setting in which an agent, embodied in a robot, learns from a human teacher. Along with a description of this framework, the most important parameters were discussed.

It is then described how the combination of these two frameworks are used to explore the questions formulated in section 1.6.

Chapter 4

Experiments in simulation

This chapter reports on simulated experiments which are based on the model described in chapter 3 and which serve as a benchmark. A baseline experiment is provided to allow for a comparison with variations of concept learning. An alternative method of teaching and testing, dubbed direct instruction (DI), is compared to the standard model of language game interaction. Next, different perceptual bases that serve to partition the CS are explored; the normal means of representing categories in a CS is compared to other methods based on k-Means and self-organising maps (section 2.4.4). Furthermore, the ability to represent prototypes in a CS representation, a property that is deemed important for concept modelling (section 2.3.5), is investigated.

4.1 Baseline

A guessing game simulation (teacher-learner scenario) was run to serve as a baseline experiment. In this simulation one agent is adopting the role of teacher and as such is endowed with categories and word labels at the beginning of the simulation. The other agent (learner) starts with an empty repertoire of categories and word labels; through interaction the learner gradually acquires categories that match those of the teacher. We measure the communicative success of both agents; recall that in the methods adopted in this work communicative success can never reach 100%, because it is averaged over the total of interactions (section 3.2.6). This simulation

was run with abstract training data, in the sense that objects in the context do not represent any particular domain but rather consist of abstract normalised vectors. The interaction was replicated 25 times ($R = 25$), to allow for statistical analysis. For every replica the number of dimensions was randomly determined through a normal distribution with a mean of 3 and an SD of 2[1]. Also the conceptual knowledge of the teacher was randomised; the number of concepts was determined through a normal distribution with a mean of 20 and and SD of 2. Other parameters were: $N_I = 2000$, $C = 4$ and $D_{min} = 0.1$. These parameters were chosen based on pragmatic considerations; they are believed to represent a reasonable 'default' setting for language game learning and as such are suitable for a baseline simulation. The resulting learning curve is shown in figure 4.1. This result is used to compare the modifications of language games that are subsequently explored in section 4.2.

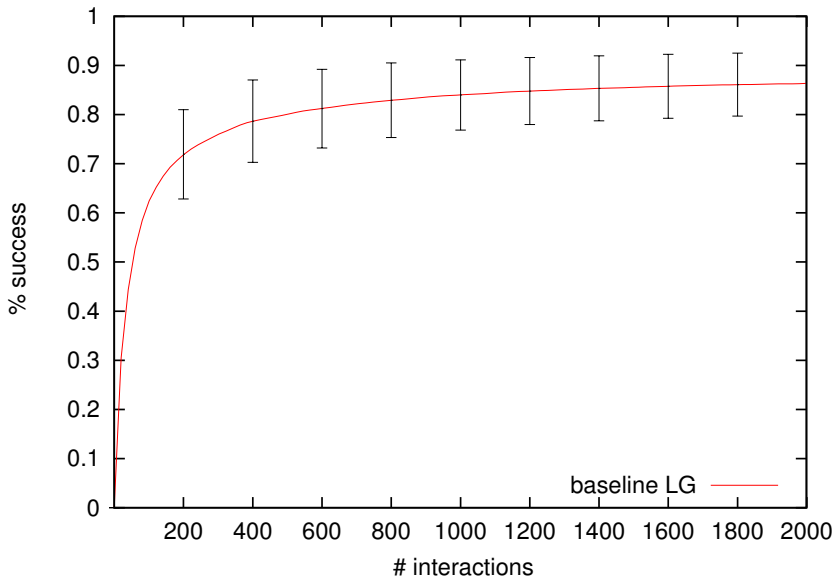


Figure 4.1: Baseline performance expressed as communicative success of two agents engaged in language game using abstract training data.

4.2 Language games and direct instruction

As described in section 3.2, language game learning (LG) happens as follows. During each round, both agents examine a context which consists of objects drawn from

[1]Using these values can result in 0 dimensions. To ensure that there will always be some data, the minimum number of dimensions was set to 1.

the environment; the teacher then chooses one object from the context as the topic of 'conversation'. The word label associated with the topic is communicated to the learner. The learner tries to identify the topic in the context based on this word label and depending on the outcome of the guessing game modifies its word-concept associations.

Direct instruction (DI) can be seen as a simplification of this teaching process. In a similar vein as LG, it consist of two agents: a teacher with predefined knowledge and a learner with an empty repertoire. The teaching mechanism differs however, as during an interaction there is no context. Instead, a single random object is chosen as the topic, and the teacher provides a word label for this. The learner then adds the word label to its lexicon and object data to its CS, either by adding the object data and word label as a new concept, or by updating the best matching concept from its CS if the word label is already known to the learner.

Also, the evaluation measurement that is used to express successful interaction can be modified in a DI manner. When using (normal) LG measurement, the communicative success after each interaction is based on the ability of the learner to successfully identify the topic from the context. Alternatively, with DI measurement, a set of random sample objects (n=20) is generated, and both agents provide a word label for these objects based on their respective best matching concepts and word associations. If the word labels match, communication is a success; if they do not match, communication fails. Communicative success is then measured as the number of objects for which the word labels of both agents match, divided by the total number of objects in the set. These two methods of teaching and two methods of testing result in four different cases, as expressed in table 4.1. Simulations with these different regimes were run using the same parameter settings as the baseline learning (figure 4.1).

	LG testing	DI testing
LG teaching	$LG_{teach} - LG_{test}$	$LG_{teach} - DI_{test}$
DI teaching	$DI_{teach} - LG_{test}$	$DI_{teach} - DI_{test}$

Table 4.1: Normal (LG) and Direct Interaction (DI) learning and testing regimes.

condition	versus	t test	
$LG_{teach} - LG_{test}$	$LG_{teach} - DI_{test}$	$t(48) = 9.3457$	$p < 0.001$
	$DI_{teach} - LG_{test}$	$t(48) = -0.911$	$p = 0.367$
	$DI_{teach} - DI_{test}$	$t(48) = 1.8852$	$p = 0.065$
$DI_{teach} - LG_{test}$	$LG_{teach} - LG_{test}$	$t(48) = 0.911$	$p = 0.367$
	$LG_{teach} - DI_{test}$	$t(48) = 8.8416$	$p < 0.001$
	$DI_{teach} - DI_{test}$	$t(48) = 2.4317$	$p = 0.019$
$LG_{teach} - DI_{test}$	$LG_{teach} - LG_{test}$	$t(48) = -9.3457$	$p < 0.001$
	$DI_{teach} - LG_{test}$	$t(48) = -8.8416$	$p < 0.001$
	$DI_{teach} - DI_{test}$	$t(48) = -7.0825$	$p < 0.001$
$DI_{teach} - DI_{test}$	$LG_{teach} - LG_{test}$	$t(48) = -1.8852$	$p = 0.065$
	$LG_{teach} - DI_{test}$	$t(48) = 7.0825$	$p < 0.001$
	$DI_{teach} - LG_{test}$	$t(48) = -2.4317$	$p = 0.019$

Table 4.2: Pairwise comparisons of performance resulting from different treatments (direct instruction and language games) for both teaching and testing methods.

4.2.1 LG and DI: results

As can be observed from figure 4.2, when the testing method is LG, the performance of both LG and DI as teaching methods is very close (difference is not significant; two-sample t-test: $t(48) = -0.911, p = 0.367$). However, when DI is used as a testing method, the difference between LG and DI as teaching methods becomes much more apparent (difference is significant; two-sample t-test: $t(48) = -7.0825, p < 0.001$). All pairwise comparisons of performance are shown in table 4.2.

If we focus on DI as a testing method (regardless of the teaching method), the difference in performance can be explained in the following manner. When a context is generated for LG, the parameter minimum-distance-objects (D_{min}) ensures that there exists a minimal distance between every object in the context. D_{min} affects the number of concepts that a learning agent will form as a result of exposure to the environment. The smaller D_{min} is, the more concepts an agent will form in its CS as to be able to discriminate these small differences. In the case of testing, using a context that respects D_{min} (as happens in LG_{test}) makes the guessing task easier, compared to a direct test (DI_{test}). This is the case because for the learning agent, when confronted with a context and a word label from the teacher in the case of LG, the task is to find the topic from the context through the word label. The learning agent finds the concept in its CS that is most strongly associated with the word label and compares this to all objects in the context. The concept associated

with the word label does not need to be exactly the same for the teacher and learner in order to successfully discriminate the topic from the context, because the D_{min} ensures that other objects in the context will sufficiently differ from the topic. To illustrate this point, an example scenario is provided below.

In a 1-dimensional environment, a context may look like this: [0.22, 0.63, 0.91]. A teaching agent may have a CS populated with the following concepts: [0.2, 0.4, 0.6, 0.8]. If the first object in the context (0.22) is chosen as topic, the teacher might communicate an associated word label "word1". The learning agent may be familiar with "word1", but has it associated with a concept with value 0.28. The learner compares this concept of "word1" with the context, and will still find that the first object (0.22) is the closest match. Because the existence of D_{min}, there is less chance that other objects in the context will be associated with the communicated word, since they tend to be sufficiently different. Thus, communication may succeed even though concepts of teacher and learner differ to a certain degree.

On the other hand, when DI is the method of testing, a sample object might be generated that is marginally closer to one of the teacher's concepts than to another. The teacher will express the associated word, while in fact the sample object also matches relatively well to another concept. If the learner has a somewhat different CS than the teacher, it may be the case that another concept from the learner's CS matches best for the sample object. In this case the word associated with this other concept will not match the word provided by the teacher and the communication fails.

To experimentally test the explanation described above, a simulation with a modified DI_{test} was run. In this experiment the learner is allowed to provide both the best-matching and second-best-matching word label for a given sample object. As a result, an increase in performance can be observed (figure 4.3). This performance increase indicates that in a substantial number of cases communication fails because the given sample object resides on the border of two concepts; by allowing the second best matching word label as a valid response these cases are accounted for. DI as a testing method can actually surpass LG as testing method.

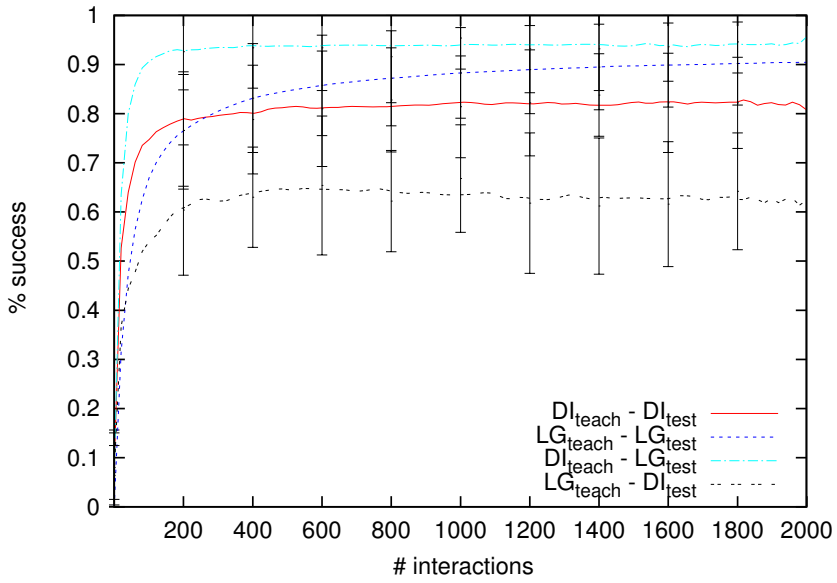


Figure 4.2: Comparison of performance under different learning and testing regimes (DI and LG).

To sum up, a test using LG gives more leeway to objects that are on the border of one concept and another, because the D_{min} ensures that other objects in the context will be sufficiently different, making it easier in those cases for communication to succeed. In other words, a test that uses DI leaves less room for successful communication about dissimilar concepts.

4.2.2 Concluding remarks

In conclusion, it appears that DI can be a useful method of teaching, because it allows for a somewhat simpler interaction scenario. However, DI as testing method is considerably harsher, and leaves less room for marginally different conceptual spaces. As such, it might not be suitable as a testing method. Also, to be able to place the research reported in this thesis in context with other research using the language game framework, we opt to retain the LG method for both teaching and testing.

4.3 Perceptual basis

When participating in a language game an agent needs to perceive the environment in order to be able to examine the context and determine a word label for the topic.

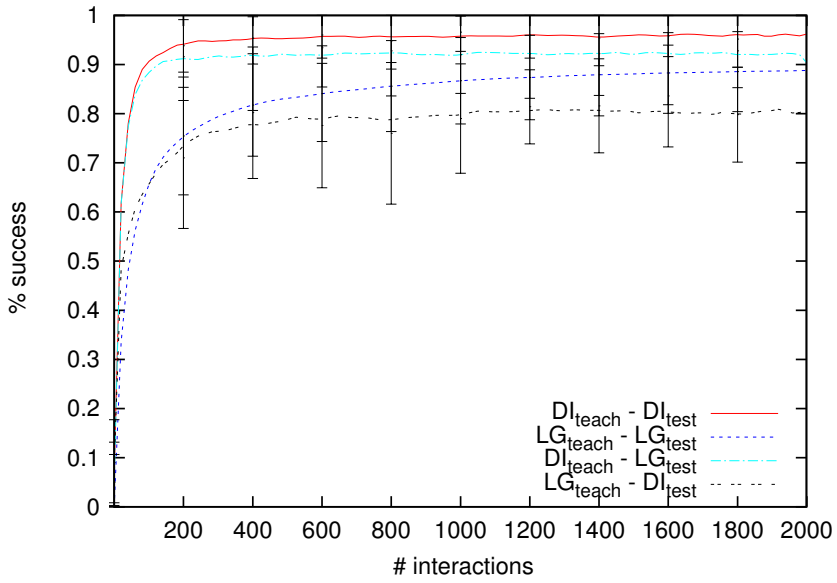


Figure 4.3: Comparison of performance under different learning and testing regimes (DI and LG), with DI utilising second best matching word label.

Therefore, agents are equipped with the perceptual means tailored towards this environment. Perception might be very direct; for instance, in a basic simulation the context typically consist of objects (e.g. colours) that are expressed as a vector (e.g. [255, 0, 0] to express a colour typically called 'red' by people). Such an encoding can directly be copied into the conceptual space of agents. In other cases perception might require some kind of pre-processing. For instance, in chapter 6 differences in perception are explored in which agents have different mapping functions to go from objective perception of the environment to subjective encoding. Or, in the case of robotic agents, camera streams might be pre-processed to extract the right kind of perceptual information.

Within the language game framework representations of the environment and the concepts that an agent learns are formed through application of discrimination games (DG, section 3.2.3). After individual agents are able to categorise new observations from the environment, a population of agents can further develop their shared meaning system by engaging in guessing games. However, to perceive the environment, application of discrimination games is not the only manner in which an agent can acquire the necessary concepts. A similar result can be achieved through the application of other clustering techniques. For instance, the well known method of k-Means clustering allows, after training, for finding clusters in multidimensional

vector spaces (section 2.4.1). Also self-organising maps (Kohonen, 1984) can be used to form a topological map (usually with fewer dimensions) of the input space (section 2.4.4). To investigate how well these techniques perform in terms of underlying representation for language games, experiments were performed in which we compared agents with three different mechanisms for categorisation of the input space: discrimination games (DG), k-Means (KM) and self-organising maps (SOM). k-Means was chosen because it is a common manner of clustering multidimensional spaces; the choice of SOMs as a means of representation was motivated by the fact that this allows for a closer comparison of the computational model with the ERA cognitive architecture (as described in section 2.4.4).

4.3.1 Perceptual basis: setup

Tests were run as follows. Training data was drawn from a normalised vector space with three dimensions; the language game parameters were $N_A = 10$, $N_I = 10000$, $C = 3$, $D_{min} = 0.3$ and $R = 25$. For each of the three mechanisms (DG, KM and SOM), categories were generated in the following fashion. In the case of DG, a standard discrimination game was used with $N = 1000$. In the case of KM, for each agent a perceptual base was generated using the k-Means algorithm from Matlab, through which 11 clusters were found from a training dataset containing 1000 normalised datapoints drawn randomly from the input space. The choice of $k = 11$ was inspired by the 11 basic colour categories as identified by Sturges and Whitfield (1995), although it is acknowledged that this is somewhat arbitrary. As the seeds on which k-Means was based were random, each agent had a different CS, albeit one that corresponded to the input space. In the case of SOM, agents were fitted with the resulting response vectors from a 4x4 SOM which was trained using the standard SOM algorithm on the input space for 1000 iterations. Thus, these agents' CSs were populated with 16 concepts. The resolution of the SOM (4x4) was based on the observation that agents fitted with a perceptual basis derived through training SOMs with a resolution of 3x3 resulted in severely less successful communication. Thus, it was reasoned that a 4x4 SOM constitutes the minimal

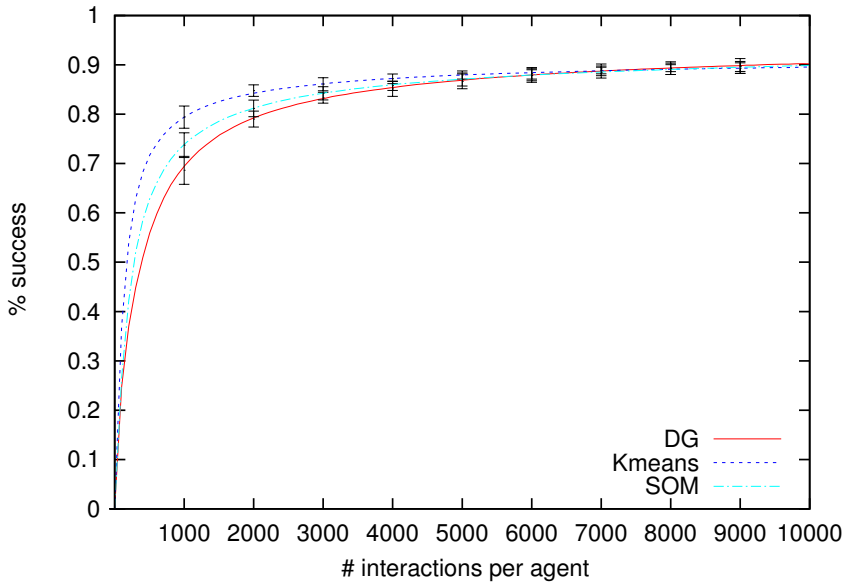


Figure 4.4: Comparison of different perceptual bases: DG, k-Means and SOM, for a population of 10 agents, where each agent interacted in 1000 guessing games. $C = 3$.

required resolution. Also in this case the seeds on which the SOMs were trained were random for each agent, ensuring that each agent had a different CS. Fitted with these different perceptual basis, agents then engaged in regular guessing games.

4.3.2 Perceptual basis: results

As can be seen in figure 4.4, the three manners of representing categories performed functionally very similar. Particularly if the simulation is allowed to run long enough, the three different regimes tend to converge. However, a main difference lies in the fact that KM and SOM need to be trained beforehand, while DG is an online learning method. For instance, for KM it is necessary to specify how many clusters we want to find, then run the algorithm and then equip agents with the resulting clusters. Also in the case of SOM, it is necessary to specify beforehand the number of nodes in the network, i.e. the resolution. Based on this, nodes in the network are shaped to reflect the input space during the training. In contrast, using discrimination games as a means to populate the CS of agents allows for a form of online learning where the agent adds new categories when necessary (as governed by the T_{adapt}, see section 3.2.7).

In a second simulation the number of objects in the context was increased to 6

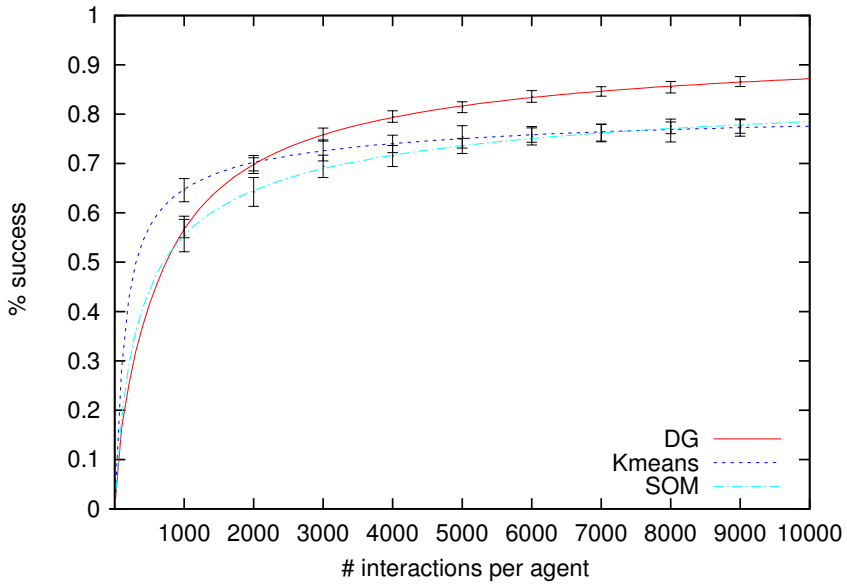


Figure 4.5: Comparison of different perceptual bases: DG, k-Means and SOM, for a population of 10 agents, where each agent interacted in 1000 guessing games. $C = 6$.

($C = 6$). This makes the guessing game harder as random guessing is only successful in 16.66% of the cases. Figure 4.5 shows the results of this regime. Interestingly, DG as perceptual base performs better with these settings. This is because DG is more adaptable; new concepts can be added as the need arises, while for KM and SOM the conceptual structures are fixed.

4.4 Modelling prototypes

This section explores the models ability to form prototypes; the work reported has been published in De Greeff et al. (2012a). As described in section 2.3.5, to account for empirical data from human concept use, a conceptual model should be able to display prototypes and typicality effects. To examine how the CS part of the model is able to build conceptual structures that exhibit these prototypes and typicality effects, the CS model was trained with data from the Zoo dataset from the UCI Machine Learning Repository (Frank and Asuncion, 2010). This database contains 101 animal exemplars which are divided into seven categories: AMPHIBIAN, BIRD, FISH, INSECT, INVERTEBRATE, MAMMAL and REPTILE. Each animal exemplar is encoded through the following 16 properties: [hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathes, venomous, fins, legs, tail, domestic,

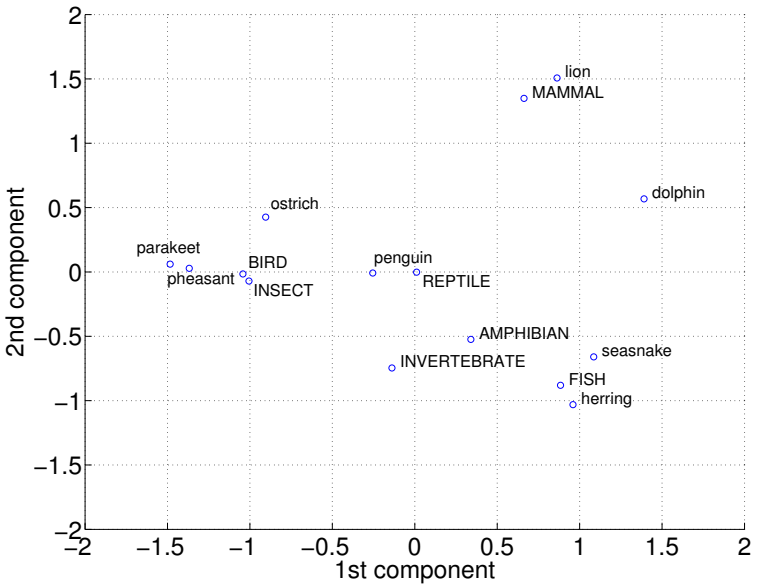


Figure 4.6: PCA showing the coordinates in the first two components for the 7 categories and the 8 animal exemplars which were used to test the CS ability to display typicality effects.

catsize]. All properties are binary, except 'legs' which is encoded as a numerical value (see appendix A for the full dataset). The 'legs' property was normalised as to allow distance measurement in the CS with the same weight setting.

From the 101 exemplars the following 8 animals were selected as test cases: dolphin, lion, herring, seasnake, ostrich, parakeet, penguin and pheasant. These 8 animal where chosen based on their explanatory value. That is, lion and herring are considered very typical for their category. Dolphin constitutes an interesting example because it is commonly incorrectly classified as a FISH, also by humans. The same holds for seasnake, as this exemplar has properties that are halfway between the FISH and REPTILE categories. The four birds are included to illustrate typicality effects within the same category. The remaining 93 exemplars were used to populate the CS in a supervised manner, following the procedure as described in section 3.1.1. Thus, prototypes were formed based on the seven categories. After training, the CS was probed with the 8 test animals. To illustrate the relation between the 7 categories and the 8 exemplars that were used to test the model's prototype structure, a principal component analysis is shown in figure 4.6.

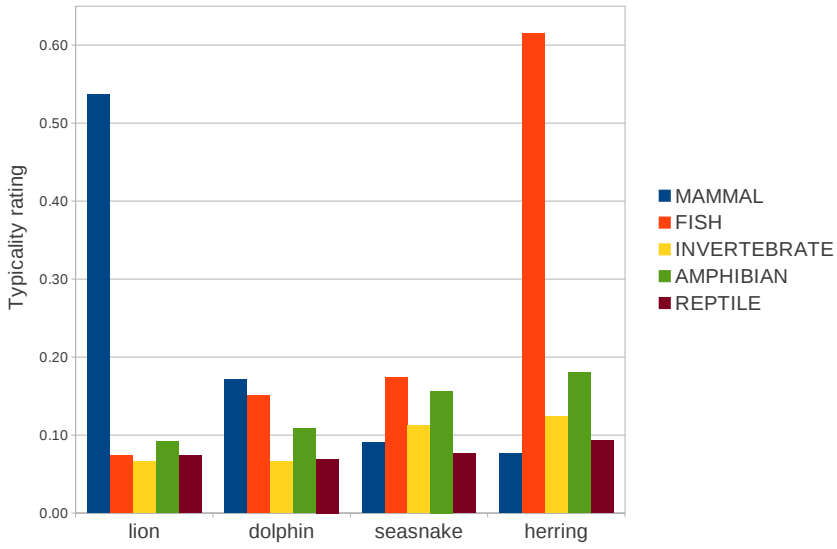


Figure 4.7: Typicality for the first four animals of the test case. All but seasnake are classified correctly; lion and herring are very typical for their category, while dolphin is atypical and very close to an incorrect category.

4.4.1 Prototype formation and typicality

Through training the CS becomes populated with animal prototypes. Next, the similarity between these prototypes and the 8 animals from the test set is calculated using equation (3.2); this results in the typicality values as shown table 4.3. The animal are classified as belonging to the category for which the typicality rating is highest. As can be observed, 7 out of 8 animals are classified correctly. Figure 4.7 shows the typicality ratings for the first four test animals. What can be observed is that both lion and herring are rated very typical for their respective categories. Seasnake is classified incorrectly as FISH (0.17), but the correct category, REPTILE, is very close with a typicality rating of 0.14. Dolphin is correctly classified as MAMMAL (0.17); interestingly FISH, which would be a common misclassification, is very close (0.15).

Focussing on the BIRD category, we can clearly observe within-category typicality effects (figure 4.8). For the BIRD class, the pheasant is the most typical example, followed by the parakeet, the ostrich and finally the penguin. This is in line with human typicality ratings (Rosch, 1975; Hampton and Gardiner, 1983; De Deyne et al., 2008), except for the fact that pheasant is rated as more typical than parakeet. Upon closer inspection it turns out that the property 'domestic', which is true for

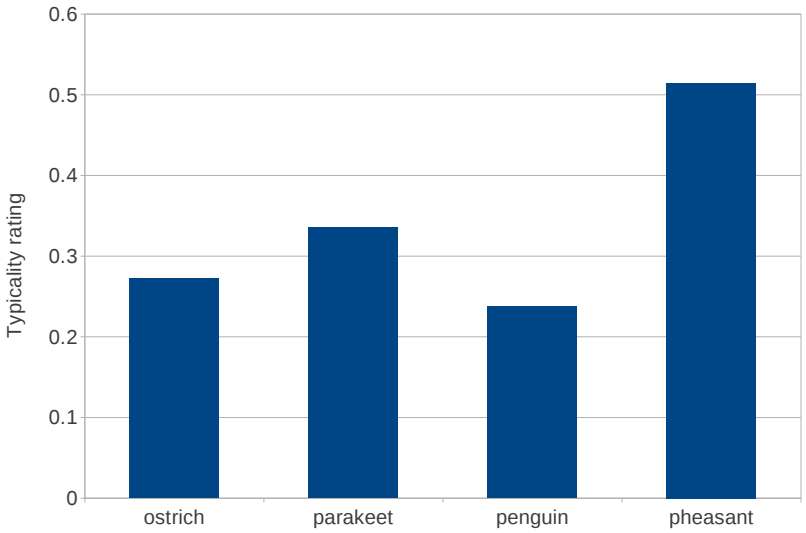


Figure 4.8: Typicality ratings of the CS model for the four bird exemplars for the BIRD category.

exemplar	classified as	INV	MAM	AMP	INS	REP	FISH	BIRD
dolphin	MAM	0.07	**0.17**	0.11	0.04	0.11	0.15	0.07
lion	MAM	0.07	**0.54**	0.09	0.06	0.14	0.07	0.07
herring	FISH	0.12	0.08	0.18	0.06	0.16	**0.62**	0.09
seasnake	FISH	0.11	0.09	0.16	0.05	0.14	**0.17**	0.08
ostrich	BIRD	0.10	0.10	0.10	0.10	0.19	0.09	**0.27**
parakeet	BIRD	0.08	0.07	0.08	0.12	0.13	0.07	**0.34**
penguin	BIRD	0.12	0.09	0.14	0.07	0.16	0.12	**0.24**
pheasant	BIRD	0.10	0.07	0.10	0.15	0.16	0.08	**0.51**

Table 4.3: Typicality ratings of the CS model for the 8 exemplars from the test set. Classifications (highest typicality rating) are shown in bold.

a parakeet, is somewhat rare for BIRD and therefore the parakeet is rated as less typical. We speculate that this contrast with typicality ratings from human data is due to the fact that the property 'domestic' may not commonly be very prominent for people when classifying birds.

Overall the results indicate that the model is able to represent concepts in a manner that is in line with the prototype theory of concepts. Detailed descriptions of this experiment can be found in De Greeff et al. (2012a) and Baxter et al. (2012). In these works we compare the ability of the model to form prototypes to the DAIM model (Baxter et al., 2011), which provides a more developmental and distributed account of cognitive modelling.

4.5 Chapter summary

This chapter has presented a selection of experiments based around the computational model as described in chapter 3. These experiments serve as an exploration of the performance of the model with respect to category learning in an interactive fashion. As there exist a large body of research that used the framework of language games (see section 3.2.1) and as such the basic dynamics are well documented, the experiments reported in this chapter constitute a somewhat 'alternative' exploration of the performance of the model. That is, rather than a systematic investigation of the impact of various parameter settings, the chapter discussed three aspects of the model that are considered important for its application in social concept learning.

Specifically, the following three series of experiments were presented. In a first series of experiments, an alternative means of learning and testing, dubbed direct instruction (DI), was compared to the classical language game methods. DI constitutes a more direct manner of interaction, effectively bypassing the use of a context (a set of objects from which agents make a guess) as is common practice in classic language games. Results indicate that the performance of the model utilising DI is comparable to that of LG. While DI as a teaching method has some merits, e.g. the fact that in the absence of a context and guessing by an agent it constitutes a more simple scenario of interactive teaching, the DI means of testing successful

communication is more harsh and as such leaves less room for marginally different conceptual spaces of individual agents. Moreover, in order to be more compatible with other work adopting the language game framework, it was decided to not use DI for further experiments.

A second series of experiments contrasted the computational model's method of learning and representing concepts through conceptual spaces and discrimination games (DG) with two other methods of partitioning multidimensional spaces: k-Means clustering and self organising maps (SOMs). The model was compared with k-Means clustering because this is a common method of classification in multidimensional spaces. The choice of comparison with SOMs was motivated by the fact that SOMs form the basis of the ERA model (described in section 2.4.4); the substitution of DG as perceptual basis with SOMs renders the computational model more comparable with the ERA architecture as it illustrates how certain functional components of both models are interchangeable. Measured by communicative success, all three perceptual bases performed very similar, with DG performing notably better in language games with a larger context. However, DG allows for online learning, as new concepts can be learned when the need arises. In contrast, k-Means and SOMs require pre-training and specification of the number of clusters for k-Means and resolution in terms of number of nodes for SOMs. As such, the DG method of learning and classification of concepts is maintained within the model.

A third experiment explored the ability of the model to represent concepts in a manner compatible with the prototype theory of concepts as discussed in section 2.3.5. The prototype theory envisions concepts as summary representations and stipulates that instances of concepts express typicality; that is, some instances are more typical for a concept than others. The ability of the model to represent concepts in this way was tested by training the model with animal exemplars from a commonly used dataset and subsequently testing it for its ability to classify and rate selected exemplars. The resulting typicality ratings were in line with empirical data from human subjects; as such the model is deemed adequate for a prototype manner of concept representation.

Chapter 5

Interactive learning

In this chapter the notion of interactive learning is discussed; rather than a passive transfer of knowledge from a teacher to a learner, a learner may influence its learning experience through active participation. Interactive learning in the context of this work is a set of modifications of the language game dynamics that allow a learning agent to be more influential with respect to what it learns. In normal language games the learner has less influence on which objects are learnt, as this is typically decided by the teacher. In contrast, by providing a learning agent with means to interact and influence the learning experience a better learning outcome can be achieved. We first discuss the relevance of interactive learning by comparing it to insights from developmental psychology, then a description of the implementation of in the computational model is given and finally improvements in learning are discussed. The work in this chapter has been published in De Greeff et al. (2009b).

5.1 Interaction

In developmental psychology it has been known for a while that social and affective interaction is central to language development and, by extension, to concept acquisition. Young learners and their caretakers engage in *intersubjectivity* (Trevarthen, 1998), the common denominator for interactions involving the learner's understanding of emotion and thought. There is evidence that the acquisition of language and specifically of vocabulary requires the young child to interact with a human

caretaker: Krcmar et al. (2007) showed how 15 to 21 month old children benefit from interaction: learning novel words was significantly more efficient when joint attention and interaction was involved. This was contrasted to learning experiences presented on a television screen, which did not involve joint attention or interaction; results showed how televised learning experiences were up to half as effective as actual interaction (interestingly, the study suggested that children are least likely to learn novel words presented by animated characters on television). Also Baldwin and Moses (2001) describe a large body of evidence indicating that children need social understanding to properly learn a word-object mapping. That is, temporal contiguity of a word with an object is often not enough for children to accept the word as a proper label, additional social cues indicating that a caregiver is indeed referring to a specific object is required to properly learn new words. These social clues may be joint attention (Baldwin, 1995; Tomasello, 1995) and/or non-verbal communication like facial expressions or gestures.

The importance of interaction for learning has also been acknowledged within the human-robot interaction community (e.g. Cassell and Tartaro, 2007). Several researchers have examined the role of interaction by building robotic models that learn from the environment through interacting with it, rather than plain observation. Breazeal and Scassellati (2000) for instance, describe a system which allows a robot to regulate the level of interaction, so that it gets neither too much nor to little stimulation from its surroundings. Thus, the system actively creates an optimal learning environment for itself.

Interaction is therefore deemed to be an important aspect, as learning is never a one-way process. As such, the computational model was augmented so that a learning agent can make use of some interactive learning features, which allow it to actively influence the learning process. These interactive features are designed to help the learning agent to develop conceptual knowledge more quickly, while keeping psychological relevance in mind. Three interactive features were developed for this end: active learning, knowledge querying and contrastive learning. Active learning gives the learner a bias to learning unknown stimuli and relates to novelty

preference which is typically observed in young children, knowledge querying helps the learning agent to solidify its conceptual knowledge by testing uncertain label-concept mappings with the teacher and contrastive learning allows the learner to use certain stimuli as negative examples, which bears similarities to lateral inhibition (Oliphant, 1999) and lexical contrast (Clark, 1993).

5.2 Adding interaction to the model

In the experiments reported here the colour domain was used as a test case, but rather than the standard RGB encoding, CIE $L^*a^*b^*$ encoding was used because this is more in line with how humans perceive colours (Fairchild, 1998). Hence, colour stimuli consisted of three values, where the L^* dimension encodes for the lightness of the colour and the a^* and b^* dimensions respectively encode for a red-green and yellow-blue dimension.

The language games framework supplied the means by which agents can learn new concepts (section 3.2). The basic functioning of this was modified to allow for more interaction through addition of components that enable the learning agent to actively steer learning experience towards gaps in its knowledge; as such this may constitute more effective concept learning. These interactive features and their implementation in language games are described below.

5.2.1 Interactive features

1. *Active learning* (AL). During a guessing game, instead of using a randomly picked topic from the context, the learner actively chooses the topic. This is done by picking the stimulus from the context for which the distance to the most nearby already learnt concepts is the greatest. That is, the most

unfamiliar stimulus is chosen as topic[1]. The idea behind AL is that selecting the most unfamiliar stimulus as the topic enables the agent to reach far corners of the conceptual space more quickly. By selecting the stimulus which bears the least resemblance to already known concepts, the agent should be able to achieve a more distributed conceptual knowledge structure. AL could be viewed as a way of modelling novelty preference which is typically observed in young children.

2. *Knowledge querying* (KQ). After a specified number of guessing games, the learner queries some of its knowledge it has learned so far with the teacher. This is done by selecting the concept which has been the least successful during previous language game interaction. This concept is presented to the teacher, along with the associated label. If the teacher confirms the query, i.e. if the label of the teacher for the queried concept is equal to the label of the learner, the strength of the association between the label and the concept of the learner is increased. If the query is not confirmed, this association is weakened. KQ aims to implement a common sense intuition, namely that it is sensible to check learned information from time to time and adapt if necessary.

3. *Contrastive learning* (CL). During a guessing game, after the learner has successfully identified the topic through the label uttered by the teacher, not only the association between label and topic is increased, but contrastive information is utilised as well. For each stimulus in the context that is not the topic, the learner finds the concept in its knowledge body that is closest, and weak-

[1]Inspiration has been drawn from Oudeyer and Delaunay (2008), which also featured a mechanism called 'active learning'. The difference with our implementation of AL and that of Oudeyer and Delaunay consists in the fact that we aim to actively explore the far corners of the conceptual space quickly. Hence, the aim is to enable the agent to experience unknown stimuli and build concepts for this. Instead, in Oudeyer and Delaunay the active selection of meaning by the agent serves as a mechanism to gradually control the growth of different meanings and thus strive for a more robust shared lexicon. Because the agent considers introducing a new meaning based on certain criteria (for instance, the average success of the meanings already in use), this active selection can be seen as a method to consolidate the knowledge already learnt, leading to faster convergence among the population. This form of AL is essentially aimed at employment within a community of agents which all interact with one another, while ours is aimed at the learning agent only. In summary, although the term "active learning" is the same, the actual implementation functions differently. It is called "active" because in both cases agents are actively engaged in the dynamics that govern the acquisition of meaning.

ens the association between this concept and the label that the teacher used to describe the topic. This is supported by experimental results from developmental psychology (O'Hanlon and Roberson, 2007) and bears resemblance to lateral inhibition (Oliphant, 1999), lexical contrast (Clark, 1993) and mutual exclusivity as used in the models described in Vogt and Haasdijk (2010)[2].

More formally, when interactive learning is used, the guessing game algorithm as specified in section 3.2.4 is modified according to the following description:

- AL. During the guessing game, when A^L is confronted with context O: (1) A^L finds best matching concept c in S_{A^L} for each stimulus in O: $\{o_1, ..., o_N\} \rightarrow C = \{c_1, ..., c_N\}$. (2) The distance between every o_i and c_i is calculated and stored in $D = \{d_1, .., d_N\}$. (3) The o_i with the highest d_i is chosen as topic for the guessing game by A^L.

- KQ. After each language game the success of the concept c^L used by A^L is recorded. After a specified number of language games A^L initiates a knowledge query: (1) A^L finds the concept in S_{A^L} with the lowest success rate c_{low} and the associated label l_{low} and communicates this to A^T. (2) A^T finds the closest concept in S_{A^T} and the associated label l_{match}. (3) If $l_{low} = l_{match}$, A^T responds positive, otherwise negative. (4) Based on the feedback from A^T, A^L increases or decreases the connection strength between c_{low} and l_{low}.

- CL. (1) After a successful guessing game A^L examines all objects $\neg o_t$ in the context and finds related $C = \{c_1, ..., c_N\}$ in S_{A^L}. (2) A^L decreases the connection between l^L and all objects in C.

5.2.2 Experimental setup

In each language game the context consisted of 4 stimuli, including the topic. This context was generated by randomly picking 4 samples ($C = 4$) from a dataset containing 25,000 pixels drawn with uniform probability from the RGB space and

[2]Mutual exclusivity as used in (Vogt and Haasdijk, 2010) however, applies to all competing concepts in an agent's knowledge body, whereas CL as implemented in this work only applies to other stimuli in a given context. As such, CL only applies to a context size ≥ 2.

converted into CIE $L^*a^*b^*$ space. Between all stimuli in the context there was a minimum distance of 50 ($D_{min} = 50$)[3]. The teacher and learner engaged in 2000 language games ($N_I = 2000$). For all learning regimes (LG, AL, KQ and CL) 300 replicas $R = 300$ were run and the average correctness score was calculated. This setup was chosen based on a reasable balance between the aim to provide a general language game learning environment within which the effects of interactive features could be explored, and the computational time needed to run the simulations.

5.2.3 Evaluation

To evaluate the performance of the different learning regimes, the conceptual knowledge held by the learner after learning sessions is compared to that of the teacher. This is done by employing a test scenario in which teacher and learner are shown a set of 100 random stimuli[4]. Both teacher and learner then state their associated label for each stimulus in the set. If the two labels are equal, the learner has learnt the label correctly. In this way the learner is assigned a correctness score S which reflects the percentage of correctly learnt labels. S is calculated as the number of stimuli correctly named by the learner divided by the total number of stimuli in the given set.

5.2.4 Result

To compare the results of the various learning regimes the LG learning method was used as a baseline performance. Then the interactive features AL, KQ and CL were compared to the baseline LG, which is shown in figures 5.1 to 5.3. Figure 5.4 displays the performance of the base condition against interactive learning with all three features enabled. The correctness score S and SD for each learning regime at the end of language game interactions is shown in table 5.1. As can be observed from the graphs, performance of interactive learning is very close to that of LG learning in all cases. CL performs slightly less, while the other learning regimes perform

[3]As an illustration of the CIE $L^*a^*b^*$ distances between typical colours: the distance between green-blue is 258, red-blue is 177, yellow-blue is 232 and yellow-green is 70.

[4]Because of running time considerations we did not use the full set of 25,000 colour samples for evaluation after each training interaction.

somewhat better than LG. With all three interactive features concurrently enabled, performance gain is the highest. To determine whether or not the differences at the end of learning were significant, a two sample t-test was performed (see table 5.2). It turns out that in all cases except for LG vs CL the difference in performance is small but significant. Overall it can be observed that learning of concepts through language game interaction may improve somewhat when interactive features are added; both in terms of learning speed and in terms of final performance after learning. The performance of any learning regime never really converges to 100% accuracy. This is due to the fact that although the learner is generally able to form concepts that functionally resemble the concepts of the teacher, the boundaries of the learners' concepts are never exactly the same. So while testing, there will always be border cases which the teacher may call 'concept 1', but the learner 'concept 2'. A pairwise comparisons of the learning performance for all different interactive features is shown in table 5.3. As can be observed, the combination of all three interactive features performs clearly the best in terms of correctness score S.

We also computed the performance of agents performing randomly, i.e. the agent picks a random label to respond to a stimulus. Typically, the performance of the agent is around 10% (this is proportional to inverse number of labels of the teacher). As can be seen from the graphs, the learners quickly perform better than random and achieve a performance of over 60% after about 100 interactions. Quite a number of concepts the agent holds at that point will in fact be learnt in just a few interactions. Hence, the learning process bears a resemblance to fast mapping in young children (Carey, 1978). The fact that the agent scores only 60% is because (1) through random selecting the stimuli not all colour categories may be encountered already at this point, and (2) the discrimination game of the teacher may fail sometimes (depending on the distance between all the stimuli in the context), rendering some of the 100 interactions not suitable for learning.

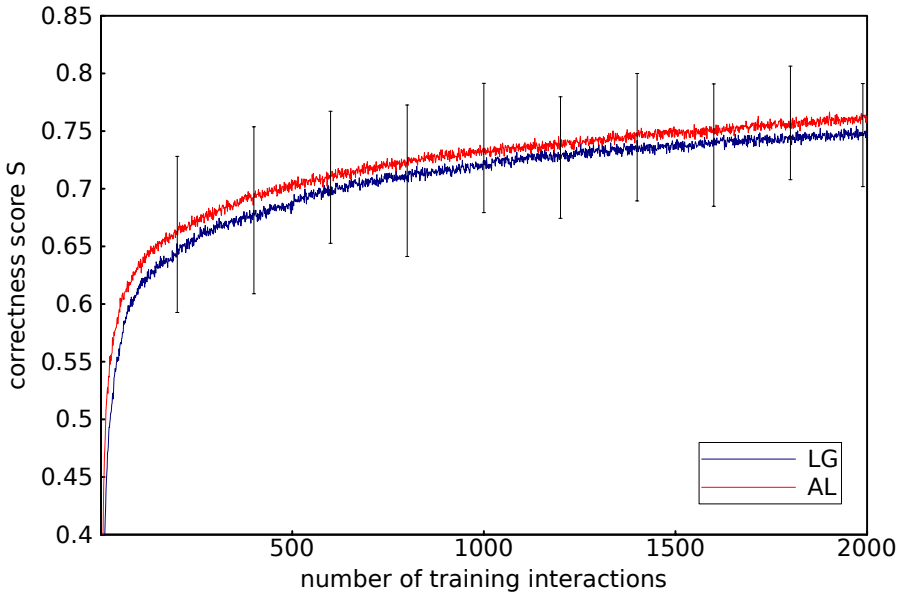


Figure 5.1: Performance of LG vs AL. The darker (blue) line indicates LG, the lighter (red) line indicates AL.

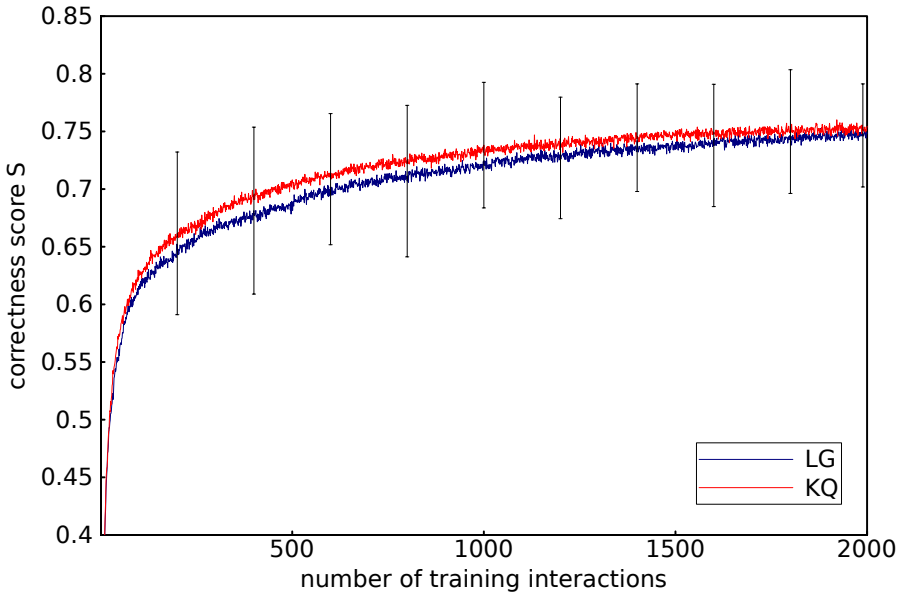


Figure 5.2: Performance of LG vs KQ. The darker (blue) line indicates LG, the lighter (red) line indicates KQ.

	LG	AL	KQ	CL	all
Average S	0.7476	0.7581	0.7579	0.7414	0.7753
SD	0.0497	0.0491	0.0515	0.0476	0.0446

Table 5.1: Average learning performance (S) and SD at the end of the language game interaction for all learning regimes.

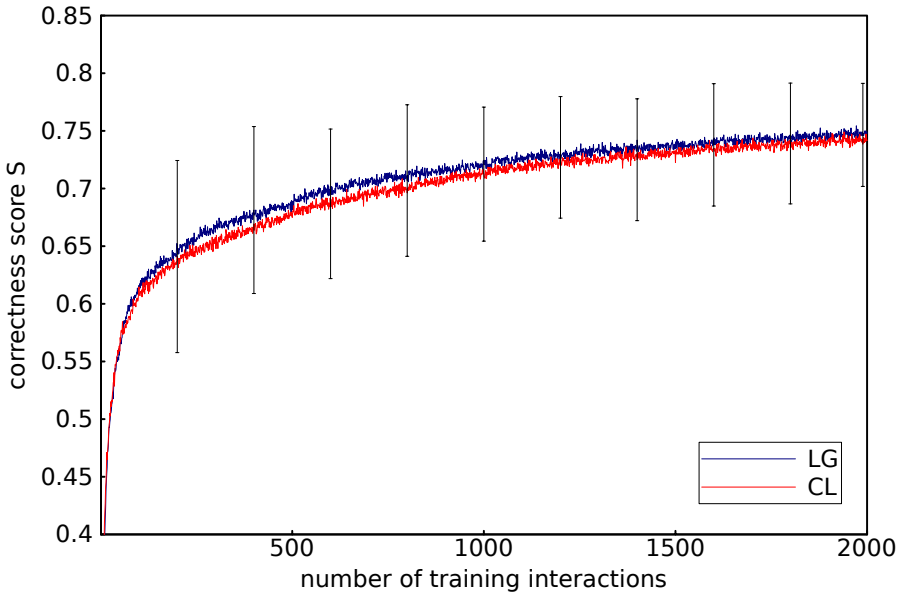


Figure 5.3: Performance of LG vs CL. The darker (blue) line indicates LG, the lighter (red) line indicates CL.

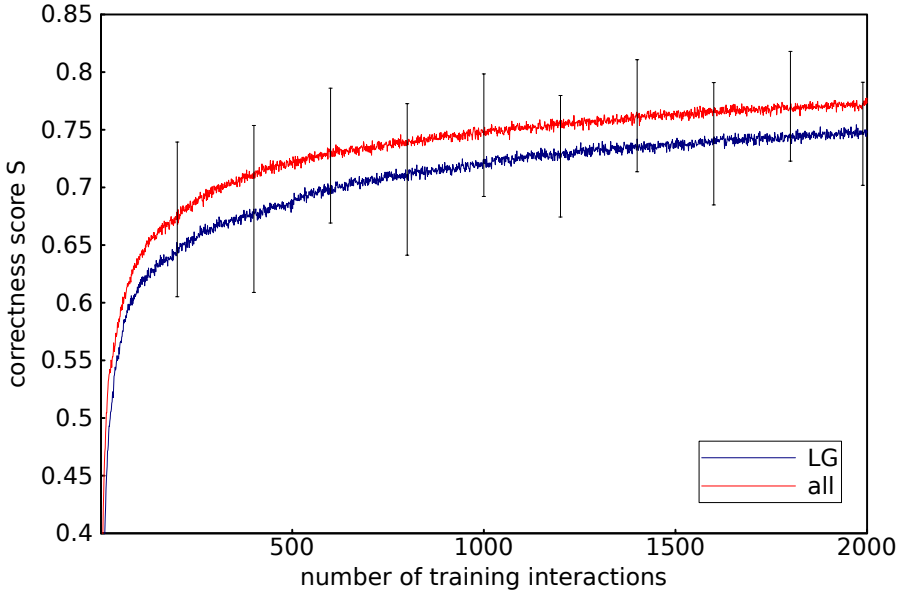


Figure 5.4: Performance of learning with all interactive features enabled. The darker (blue) line indicates LG, the lighter (red) line indicates learning with all interactive features.

	LG vs AL	LG vs KQ	LG vs CL	LG vs all
t-test	t = -2.6114	t = -2.4928	t = 1.5676	t = -7.1749
	df = 597.9	df = 597.3	df = 596.9	df = 590.9
	p = 0.0092	p = 0.013	p = 0.12	p < 0.0001

Table 5.2: Learning performance using LG compared with the learning performance using the three interactive features AL, KQ and CL, plus all three features combined.

regime	versus	t test	
AL	KQ	$t(596.7) = 0.0568$	$p = 0.9547$
	CL	$t(597.5) = 4.2472$	$p < 0.001$
	all	$t(592.6) = -4.4775$	$p < 0.001$
KQ	AL	$t(596.7) = -0.0568$	$p = 0.9547$
	CL	$t(594.5) = 4.0842$	$p < 0.001$
	all	$t(586.1) = -4.4189$	$p < 0.001$
CL	AL	$t(597.5) = -4.2472$	$p < 0.001$
	KQ	$t(594.5) = -4.0842$	$p < 0.001$
	all	$t(595.4) = -9.0015$	$p < 0.001$
all	AL	$t(592.6) = 4.4775$	$p < 0.001$
	KQ	$t(586.1) = 4.4189$	$p < 0.001$
	CL	$t(595.4) = 9.0015$	$p < 0.001$

Table 5.3: Pairwise comparisons of the learning performance for all different interactive features.

5.3 Alternative versions of active learning

The previous section illustrated how interactive features can improve learning. From the three interactive features tested AL appears to be the most effective, i.e. resulting in a small, but stable improvement. Hence, this feature was further explored.

In the experiments described in the previous section the training data that was used consisted of random points in the CIE *L*a*b** colour space. Although informative, this constitutes a somewhat abstract case. As described in section 4.4, the model can also be used to learn and represent more 'natural' prototypes, i.e. concepts that are based on a real world data and that have more properties (that is, higher dimensionality). Hence, this section explores how well AL performs when the training data is based on the Zoo database that was used for prototype formation.

As it turns out, for the Zoo dataset, the addition of AL is less effective; particularly during the earlier interactions (< 1000) the addition of AL results in a slightly lower communicative success (compare NO AL to AL1 in figure 5.5). This sparked the investigation of alternative implementations of AL. A modification that performed better was implemented as follows: when a learning agent observes a given context to determine what is the least known item, it calculates for each item in the context the distance to all categories in the agent's repertoire. In the experiments described in the previous section, the distance to the *best* matching category was stored and subsequently the item with the largest distance was selected as topic

(this 'classic' version of AL will be referred to as AL1 henceforth). Here, the distance to the *least* matching category is stored for each item (this version of AL will be referred to as AL2). Then, the item that has the highest distance becomes the learning preference for the agent. A description of AL2 using the same format as section 5.2.1 is provided here:

- AL2. During the guessing game, when A^L is confronted with context O: (1) A^L finds the *least* matching concept c in S_{A^L} for each stimulus in O: $\{o_1, ..., o_N\} \rightarrow C = \{c_1, ..., c_N\}$, by calculating the distance between o_x and each c in S_{A^L} and selecting c for which the distance is greatest. (2) The distance for every (o_i, c_i) pair is calculated and stored in $D = \{d_1, .., d_N\}$. (3) The o_i with the highest d_i is chosen as topic for the guessing game by A^L.

The rationale behind this alteration is that by selecting the least matching category for each item in the context and subsequently choosing as the topic the item for which this least matching distance is the greatest, AL2 constitutes a more 'extreme' version of AL and as such results in a more effective use of the space. The simulation was run with the following parameters: $N_I = 2000$, $C = 3$ and $R = 25$; the notion of D_{min} was not used, as the training examples were not generated datapoints, but animals from the Zoo dataset. As can be observed in figure 5.5, AL2 is particularly effective for training data with high dimensionality, as is the case for the Zoo dataset. Note that although AL1 is less effective, on the long run it still outperforms the base condition (NO AL). The difference in performance between NO AL and both AL1 and AL2 is significant, with $t(48) = -11.3535$, $p < .0001$ and $t(48) = -35.9831$, $p < .0001$ respectively (two-sample t-test).

5.4 Discussion

As described above, the addition of interactive features to language game learning can improve the effectiveness of the algorithm. Different interactive features have different effects: knowledge querying marginally improves learning, contrastive learning does not constitute any improvement (the difference compared to a normal

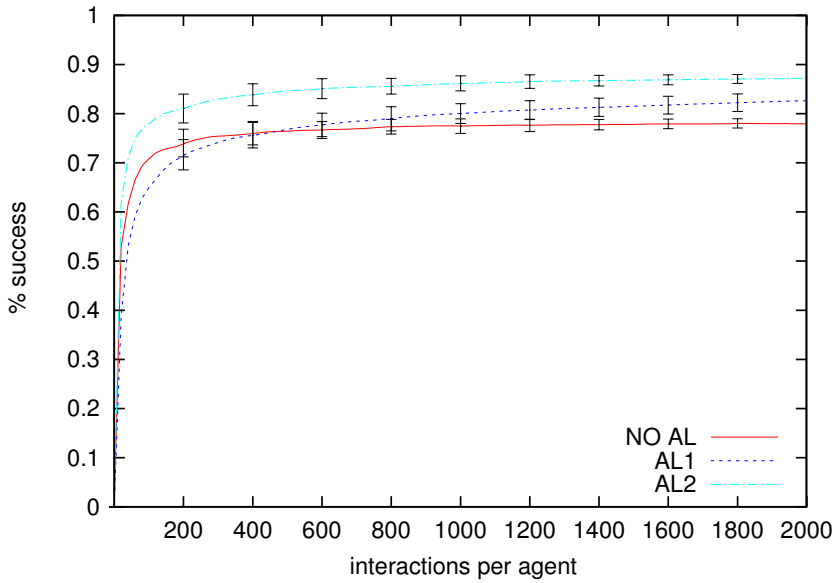


Figure 5.5: Performance of normal learning (NO AL) contrasted with forms of active learning (AL1 and AL2). Both versions of AL outperform NO AL on the long run.

language game is not significant) and active learning actually improves the learning. While the differences compared to a setup that does not utilise interactive features is relatively small, it nevertheless constitutes a significant improvement and as such deemed useful. Furthermore, particularly active learning illustrates how a learning agent can benefit from actively influencing its learning experience. Hence, active learning will be further put to use in section 7.4, in which a robot employs this strategy while learning from a human tutor.

5.5 Chapter summary

This chapter has discussed the notion of interactivity during learning. Various studies haves show that learning for people is very often an interactive process, in which both teacher and learner contribute to the learning experience. Given this importance of interactivity, this chapter has illustrated how the language game framework as a learning algorithm can be augmented with some interactive features. Particularly the feature of active learning, which results in the learner being able to more effectively populate its conceptual space, results in a small, but significant improvement. An alternative version of active learning was also explored, which turned out to be more effective for datasets with higher dimensionality.

Chapter 6

Difference in embodiment

This chapter discusses the effect of different embodiments and different perception of the world on the development of shared meanings between agents. Using colour as a case study, we discuss how despite perceptual differences agents can develop a common understanding of colour categories. This phenomenon is investigated through computational modelling of agents with different perceptual capabilities that engage in linguistic interaction. Difference in perception is modelled on both human physiological differences and on data recorded from two types of robots.

We investigate the notion of category alignment with perceptual differences through two experiments. In the first experiment artificial agents are endowed with an 'individualised' manner in which they perceive the environment; their performance in language games is compared to agents with normal perception. In a similar fashion performance in language games from agents whose perception is embodied in two different robots is compared to those of agents with normal perception. In a second series of experiments the agents' difference in perception is more closely modelled on human physiological differences in perception; again this is compared to normal perception. The work in this chapter has been published in De Greeff and Belpaeme (2011a,b).

6.1 Difference in embodiment and perception

Young children typically learn new categories through interaction with their caregivers. As discussed in section 2.3.1, it is well known that language is hugely important in this process, as the specific words that are used influence the categories a child will form. Drawing on this psychological data, models have been created to simulate the development of new categories within a population of artificial agents (e.g. Steels and Belpaeme, 2005). In these models, simulated agents perceive an environment and develop new categories through linguistic interaction. Typically the perceptual capabilities are the same for all agents in the population, i.e. the agents are homogeneous.

As discussed in section 1.2.1, cognition (for both humans and artificial agents) is considered to be embodied. Rather then merely 'facilitating' an environment, the physical body actually shapes cognitive processes, which implies that variation in the human body and specifically variation in the sensor modalities will result in varying perception. When the perception of two agents is not identical this could lead to diverging perceptual categories and concepts. One would expect this to have a negative impact on communication: if two agents have categories and concepts that are not the same, then communication is expected to be affected. Also the learning of meaning from one agent to another would be expected to be influenced by perceptual differences.

There are striking differences to be found in the physiology of human perception (Roorda and Williams, 1999). This is not only the case between adults, but also between caregivers and children. A child with a developing body and neural system perceives the world differently throughout different developmental stages (Ling and Dain, 2008). More specifically, regarding human perception, it has been shown that the ratio of cones sensitive to medium wavelengths (M cones) and long wavelengths (L cones) in the human retina can be very different between individuals. The different number of L and M cones should, according to neurophysiological understanding of colour perception, lead to broad variations in colour perception. However, people with different L:M cone ratios do not seem to vary much in colour perception. When

asked to point out the colour on a spectrum they see as unique yellow, they report virtually identical wavelengths. To explain this phenomenon, it was suggested that "neural factors play an important role in stabilizing unique yellow against variation in the L/M cone ratio" (Brainard et al., 2000).

What these "neural factors" might be is still unclear. Various plasticity mechanisms have been proposed which allow for tuning colour experience in the higher layers of visual perception (Neitz et al., 2002). However, Solomon and Lennie write "In the retina we still know little about [...] why human colour vision seems to be hardly affected by variation in the proportions of cones of different types." (Solomon and Lennie, 2007, p. 284). Hence, it is still unclear by which process the perceptual experience of colour becomes attuned, so that individuals experience colours in a similar manner despite having individual neurophysiological differences.

While neural factors certainly might play a role, an alternative hypothesis is that language might be the mechanism responsible for coordinating subjective colour experience. It has been suggested that language plays a role in the apparent shift of the lateralization of colour perception in infants and adults (Franklin et al., 2008). Thus, this would indicate that language influences the developmental trajectory of colour perception. This could also explain why colour deficiencies go relatively unnoticed. In protan colour deficiency the L cones are missing or abnormal, while in deutan colour deficiency this is the case for the M cones. About 1% of Caucasian males has protanomaly and 6% has deuteranomaly (Sharpe et al., 1999). However, colour deficiencies are often picked up rather late and only through the administration of a colour deficiency test (such as the Ishihara colour blindness test; Ishihara, 2001). This suggest that other processes might be at work which moderate perceptual differences caused by colour deficiencies.

In this chapter we explore an alternative explanation for why the differences in cone ratios does not cause people to perceive colours very differently. We show how linguistic mechanisms allow agents to align their colour perception with respect to their environment. That is to say, through linguistic interaction with others, people may be able to reach an agreement on which perceptual experiences are to

be called red, green, yellow, etc. in such a way that allows for variations at the lower neurophysiological level. We show the feasibility of this position by presenting a series of computational models in which populations of agents are endowed with varying perceptual abilities, but are nevertheless able to achieve a shared system of conceptualisations with respect to colour terms. In other words, the individual differences that exist in lower level perception do not hinder effective communication because of a dynamic process of shaping linguistic meanings through interaction with other individuals.

6.1.1 Human colour vision

The neurophysiology and psychology of colour perception has been well studied (Wyszecki and Stiles, 1982; Gegenfurtner and Sharpe, 1999), making colour an ideal test ground for cognitive models. Humans have four types of photosensitive receptors in the retina. The achromatic rod receptor contributes little to colour perception and mainly serves scotopic vision. Colour is perceived by three types of chromatic receptors, known as cone receptors. Each of the three cone types has a different peak sensitivity: to short wavelengths (S-cones, 430 nm), to medium wavelength (M-cones, 530 nm) and to long wavelengths (L-cones, 560 nm), but respond to a wide range of visual light. The sensitivity curve of a specific type of cone is an indication of the probability that this type will absorb a photon of a particular wavelength (Solomon and Lennie, 2007); see figure 6.1 for a plot of the sensitivity curves for the three cone types (Stockman and Sharpe, 2000). Hence, the human colour vision is trichromatic. In trichromatic colour vision the perception of colour depends on the combination of activation of three different cone types.

6.1.2 Physiological differences

Regarding the organisation of the retinal mosaic of the three cone types, it has been quite well established that the number of S-cones is about 5 to 10% and that the distribution pattern is semi-regular (Williams et al., 1981b,a; Hofer et al., 2005). However, with respect to the M and L cones, there appears to be a much wider

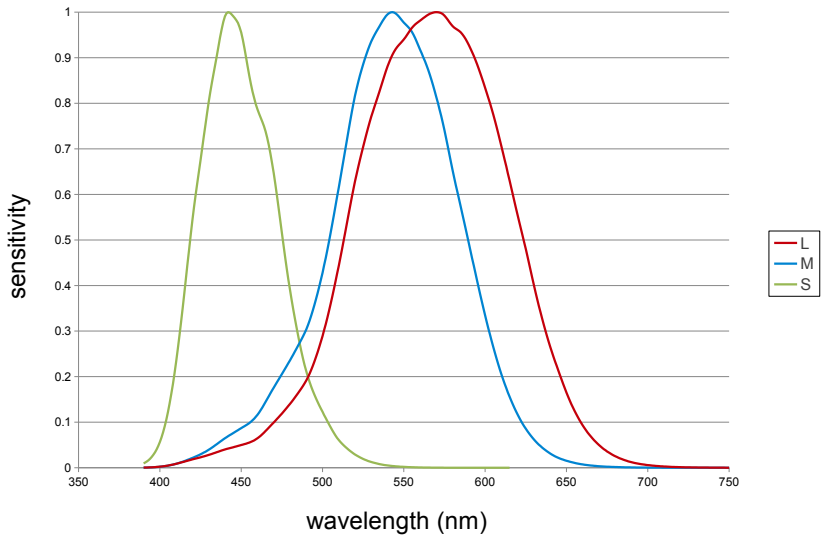


Figure 6.1: Sensitivity curves for the three human colour cone receptor types (2-degree fundamentals from Stockman and Sharpe, 2000)

variation, both in the relative proportions of M and L cones and with the distributional patterns. Contrary to what was assumed, the amount of M and L cones to be found in the retina is not roughly equal for individual humans. Big differences between individuals can be found upon close examination. Roorda and Williams (1999) reported a L:M cone ratio of 1.15 for one subject and 3.79 for another; see figure 6.2 for an illustration of the trichromatic cone mosaic in pseudo-colour for two subjects. In another series of experiments 62 normal males were tested and substantial individual differences were observed, with L:M cone ratios ranging from 0.4 to 13 (Carroll et al., 2002).

Brainard et al. (2000) studied functional consequences of the relative numbers of L and M cones. In particular, they focussed on the effect of a varying L:M ratio on the perception of unique yellow, as yellow light is picked up by both L and M cones. As the ratio between L and M cones was 1.15 for one subject and 3.79 for the other, it was predicted this would have a large effect on the wavelength of light that both subjects perceived as unique yellow. For this prediction they used an additive model proposed by Cicerone (1987) which assumes that the contribution of L and M cones depends on their relative numbers and which predicts how the wavelength of unique yellow will vary depending on different L:M cone ratios (equation (6.1)).

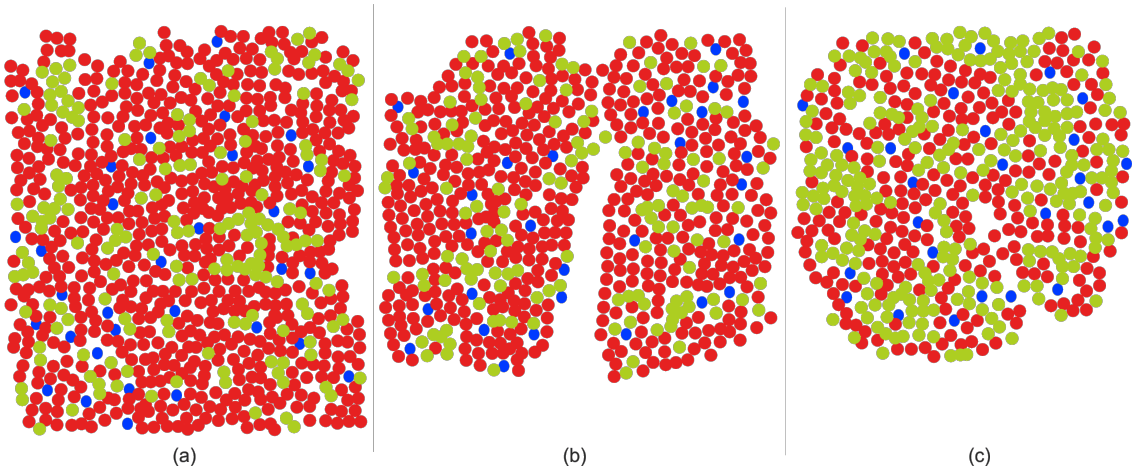


Figure 6.2: Image of the trichromatic cone mosaic in pseudo-colour for one subject (a and b) and another (c), adapted from Roorda and Williams (1999). Blue, green and red are an indication for S, M and L cones respectively.

$$(N_l/N_m)L(\lambda_y) - kM(\lambda_y) = 0 \tag{6.1}$$

where λ_y is the wavelength specifying unique yellow, N_l/N_m is the L:M cone ratio, k a parameter which tunes the relative contribution of M and L cones to the subsequent red-green channel and $L(\lambda_y)$ and $M(\lambda_y)$ express the spectral sensitivity of the L and M cones respectively. This model predicts how the wavelength of what individuals perceive as unique yellow varies as a function of their L:M cone ratio. In a later study this was empirically confirmed (Otake and Cicerone, 2000).

However, Brainard et al. found that the variation in individual L:M cone ratios was large, while the variation in their judgement of unique yellow varied only slightly (with two subject reporting 576.8 and 574.7 nm). According to the contribution of L and M cones to the perception of yellow, individual variation in unique yellow should have spanned a range between 500 and 600 nm, but this was not observed.

6.1.3 Neural factors

Neitz et al. (2002) offered an account of these neural factors: "a neural normalization mechanism for colour perception, determined by visual experience, operates to compensate for large genetic differences in retinal architecture and for changes in chromatic environment." They argued for the existence of such mechanism by

having subjects wear coloured lenses for certain periods of the day and subsequently finding differences in reported wavelengths of unique hues, both when perceptual alterations were present and not present. As such, they concluded that "the global normalization described here apparently compensates for the huge genetic variation in the ratio of L to M cones."

However, the individuals who were subjected to these perceptual alterations, were adults with normal colour vision. As such, they had a lifetime of experience with normal colour perception, which was most likely internalised in some kind of colour mapping. Hence, upon experiencing a shift in perceived wavelengths when observing objects and/or scenes from which they would know which colours are appropriate through previous experience, it is only natural to assume that this previous knowledge allows for a consistent shift in the whole colour experience. In other words, it may be very likely that a subject who knows that a certain object is yellow and who experiences the perception of this object through altered perception, could come to regard the shifted wavelength of the object as yellow again after some prolonged exposure. This would be the case in particular when the subject would also engage in interaction with his/her surroundings about the colour of objects, which is to be expected when people are subjected to this kind of perceptual alterations. So, although Neitz et al. showed the existence of plasticity on the perceptual level, this is not necessarily an explanation of how colour perception became aligned in the first place for individuals with different cone ratios.

6.1.4 Linguistic factors

An alternative explanation for the close agreement between individuals on colour terms might be of a cultural rather than neural account. That is, through a shared language people with significantly different colour perception might be able to negotiate colour terms and come to a mutual agreement. Neitz et. al. dismissed this notion by arguing that "we show ... that anomalous observers give different values for unique yellow compared to color normals, even though they are subject to the same cultural influences." (p. 788).

However, again, the apparent existence of this plasticity mechanism which operates on a relative short timescale (the shift in colour perception persisted for 1 to 2 weeks after the filters were discontinued) does not necessarily rule out any cultural and linguistic factors. Linguistic factors operate on a much more prolonged timescale and would predominantly be active throughout language development, i.e. during childhood. Indeed, different studies have demonstrated that language has an impact on perception (Boroditsky, 2006; Tan et al., 2008) and also on colour vision (Gilbert et al., 2006; Regier and Kay, 2009). As such, we will further explore how through linguistic mechanisms a group of individuals with different perception can come to agree on colour terms.

6.1.5 Computational experiments

As there is variation in the distribution of photosensitive receptors in the human retina and variation in how people name colours, we aim to capture this 'individual' perception and linguistic production in a computational model. The basics of the model operate according to the description provided in section 3.2; the way in which agents perceive their environment is modified as to capture individual differences due to embodiment. Two series of experiments are described: experiment 1 concerns a comparison of normal agents with another group with altered perception based on robotic data; experiment 2 provides a similar comparison, but the individualised perception of agents is more closely modelled on the human mechanism of colour vision and a more extensive analysis is provided.

6.2 Experiment 1

6.2.1 Synthetic experiments with agents with individualised perception

Typically, in language games, interacting agents have the same means of observing their environment. That is, participating agents perceive the context in exactly the same manner (i.e. values of the properties of the scene are extracted the same way

for all agents). In contrast, this is not the case for humans. As described above, individual humans may vary greatly with respect to their cone distribution and yet describe the same colour stimulus in the same terms. To simulate this effect, the perceptual capabilities of artificial agents were modified in order to reflect such differences in perception. In this first experiment, it was assumed that the differences between individuals in terms of L/M cone ratio is arbitrary. To model this the perceptual function of agents was altered by systematically modifying their observations with a random (but persistent within an individual) factor. This resulted in a unique manner of perceiving the environment for each individual agent, reflecting the difference in human individuals. We compared the success in communication of agents with these modified perceptional capabilities to a group of agents that had normal perceptual capabilities.

During a language game, when data from the context was processed by the agents, the RGB values were modified on an individual basis for each agent according to equation (6.2):

$$s^a = f(s_{rgb}) \cdot W^a \tag{6.2}$$

where s^a is the stimulus as received by agent a, f is a function converting data from RGB to LMS colour space, s_{rgb} is the unmodified stimulus and W^a is a set of weights specific to agent a. In this way each agent perceived a 'personalised' version of a given stimulus (appendix B provides a detailed description of this conversion). Figure 6.3 shows a typical example of the projection of 11 basic colour coordinates (Sturges and Whitfield, 1995) from RGB to an agent's LMS space.

This setup was tested in two different manners. The first was a scenario in which one agent acted as a teacher and thus had predefined knowledge of 11 basic colour categories which were learned by another agent with an empty conceptual repertoire. The second manner was a scenario in which a population of agents all started with an empty conceptual space and gradually developed a shared system of meanings by altering teacher and learner roles over the course of development. Parameters were $N_I = 1000$, $C = 3$, $D_{min} = 0.3$ and $R = 100$ for the teacher-learner

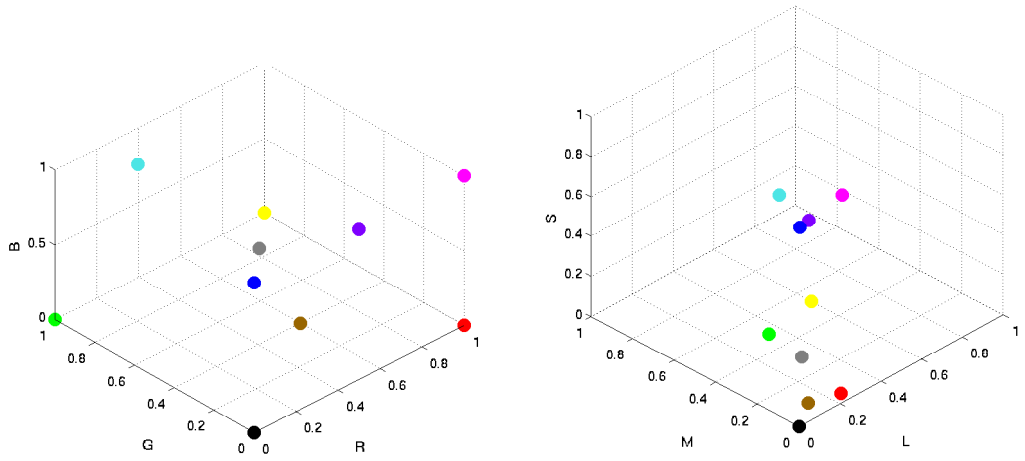


Figure 6.3: Projection from RGB to LMS colour space.

scenario, and $N_A = 20$ and $N_I = 5000$ in the population scenario. Parameter choise was based on pragmatic considerations (such 'normal' language game development as computational time).

6.2.1.1 Teacher-learner scenario

Experimental setup Sets of contexts consisting of 3 random RGB stimuli were generated. One agent was endowed with conceptual knowledge of colour terms, i.e. its conceptual space was populated with 11 basic colour terms and their coordinates, translated to the agent's LMS encoding. This agent functioned as a teacher and the other agent as a learner. One series of language game interaction consisted of 1000 language games and this setup was replicated 100 times to obtain a measure of average communicative success.

Results The performance of the group with individual perception compared to a group of agents with normal perception is shown in figure 6.4. As can be seen, agents with individual perception communicate less effective compared to the performance of agents with normal perception (two-sample t-test with $t(198) = -62.7997$, $p < 0.0001$). Nevertheless, the agents are able to reach a communicative success of around 80%, compared to just above 90% for the agents with normal perception. Also, the shape of the performance curve is very similar over the course of development.

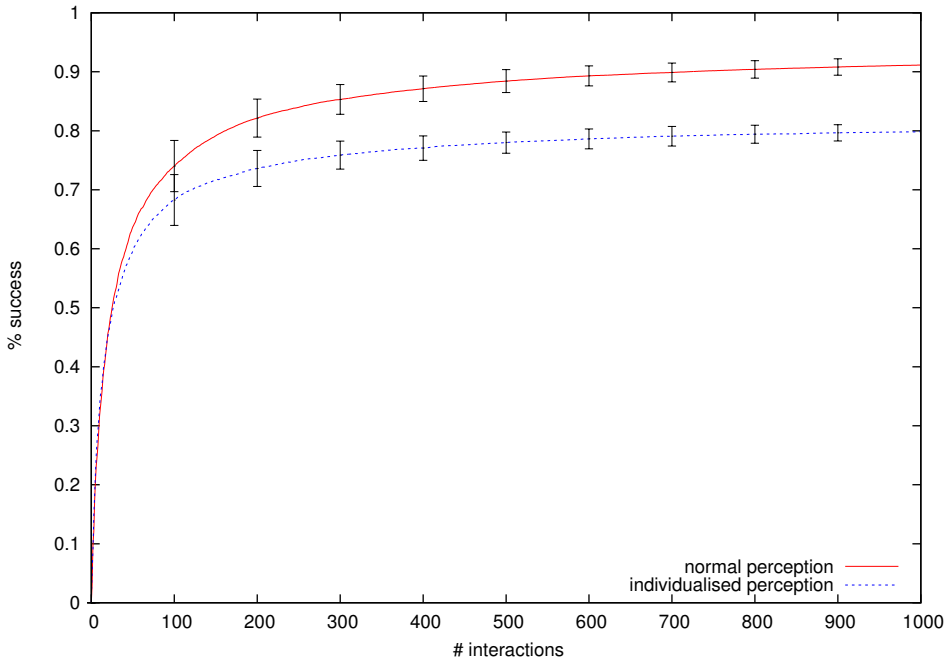


Figure 6.4: Performance of agents with normal perception compared to performance of agents with individual perception (error bars show SD).

6.2.1.2 Population scenario

Experimental setup In addition, another experiment was run with the same parameter settings, but instead of only a teaching and a learning agent, a population of 20 agents engaged into language games. For each interaction, two agents were selected randomly to act as teacher and learner. All agents in the population started with an empty conceptual space, so no knowledge of colour concepts was predefined. During a series of interaction 50,000 language games were played. The setup was replicated 100 times.

Results On a population level, the difference in performance becomes smaller. The communicative success of the population of agents with individual perception is rather close to the performance of the population with normal perception (figure 6.5). However, the difference is still significant (two-sample t-test with $t(198) = -20.1384$, $p < 0.0001$). These results are discussed in more detail in section 6.2.3.

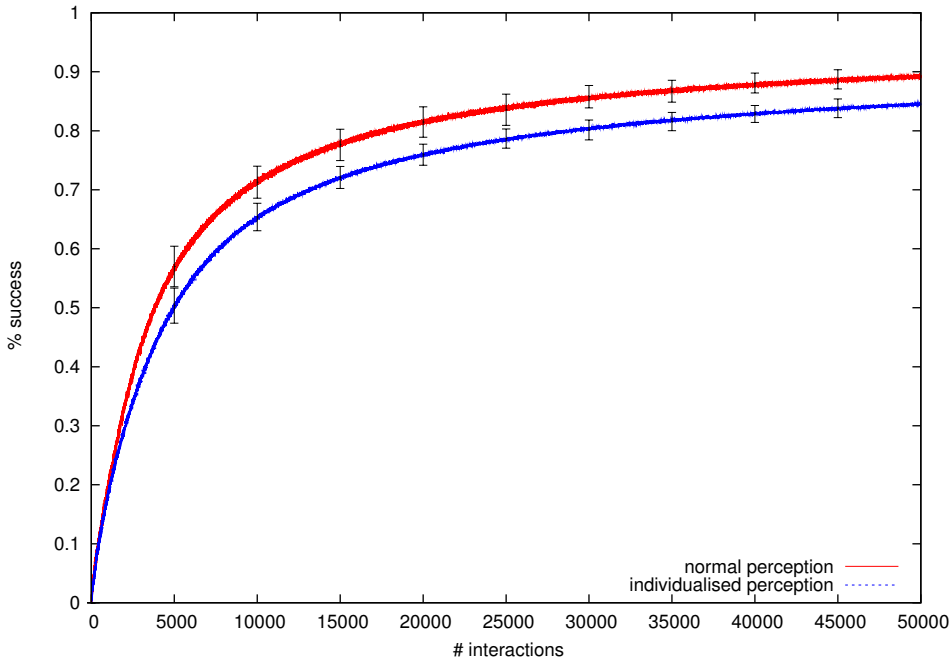


Figure 6.5: Performance of a population of agents with normal perception compared to a population of agents with individual perception (error bars show SD).

6.2.2 Synthetic experiments with data recorded from embodied robots

In this experiment, data for stimuli was recorded by having two robots examining the same context. The two robots each perceive differently, not only because of their physiological make-up, but also because of their unique perspective on the scene which resulted in different lighting conditions. One robot was an iCub humanoid robot (Metta et al., 2008) and the other the LightHead robot, which is a newly developed robot specifically tailored for HRI. The LightHead robot is described in more detail in section 7.2. Both robots are quite different in terms of hardware (CCD sensor for the iCub cameras, CMOS for LightHead's), capabilities and cost. Figure 6.6 shows an example of different perception of the same stimulus, which led the iCub robot to encode the colour as RGB(220, 124, 85) and the LightHead robot to encode the colour as RGB(178, 49, 21). Even though the colours appear to be rather similar, they constitute a considerable distance in RGB space (figure 6.7).

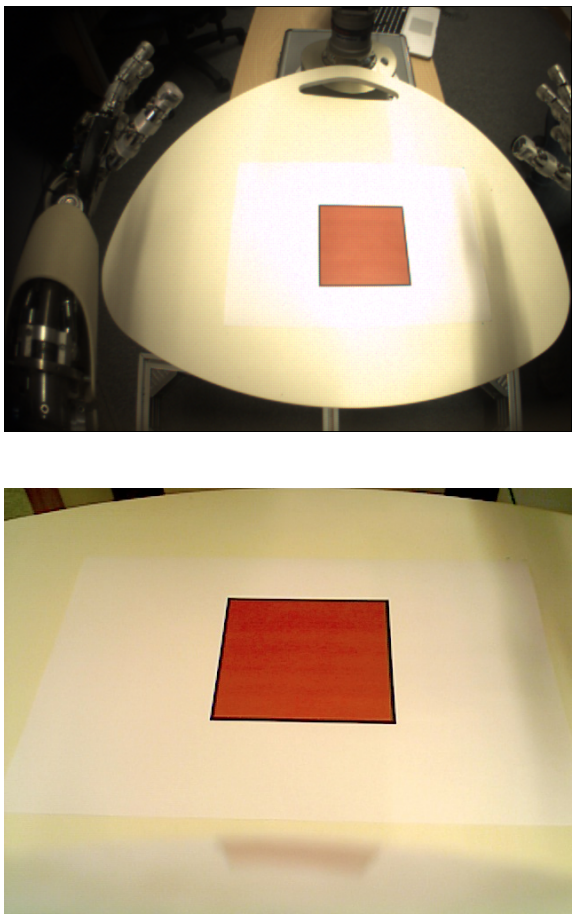


Figure 6.6: Perception of the same stimulus by the iCub robot (top) and the LightHead robot (bottom).

6.2.2.1 Data collection

Data was recorded by having both robots observe a scene in which colour stimuli were presented (figure 6.8). Colour stimuli consisted of the 11 basic colour centroids as reported by Sturges and Whitfield (1995) with an additional 39 colours randomly generated from RGB space. Both robots took a snapshot of the stimuli and saved it for off-line processing. In this manner both robots recorded 50 stimuli. After recording, the RGB values were extracted by running a blob-detection algorithm to find the precise location of the stimuli within the recorded picture and calculating the average RGB values for this region. The RGB data was not converted to LMS because this colour space models human cone reception. This procedure resulted in a set of 50 colour stimuli for the iCub robot and 50 stimuli for the LightHead robot. This set of colour stimuli was then used to play language games with artificial agents. Parameters were the same as the experiments described in section 6.2.1.

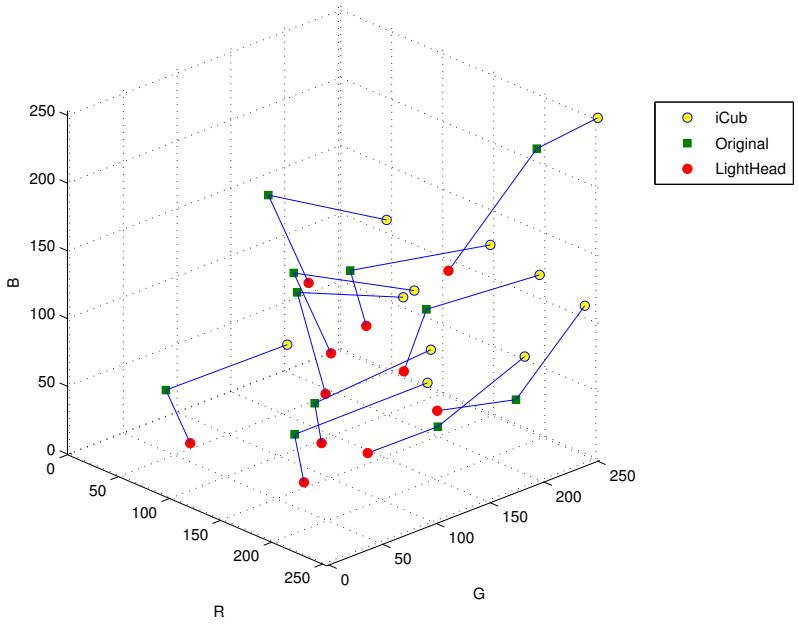


Figure 6.7: The RGB values of 11 colour stimuli plotted on their original position, and the positions as observed by the iCub and LightHead robot. Matching colours are connected.

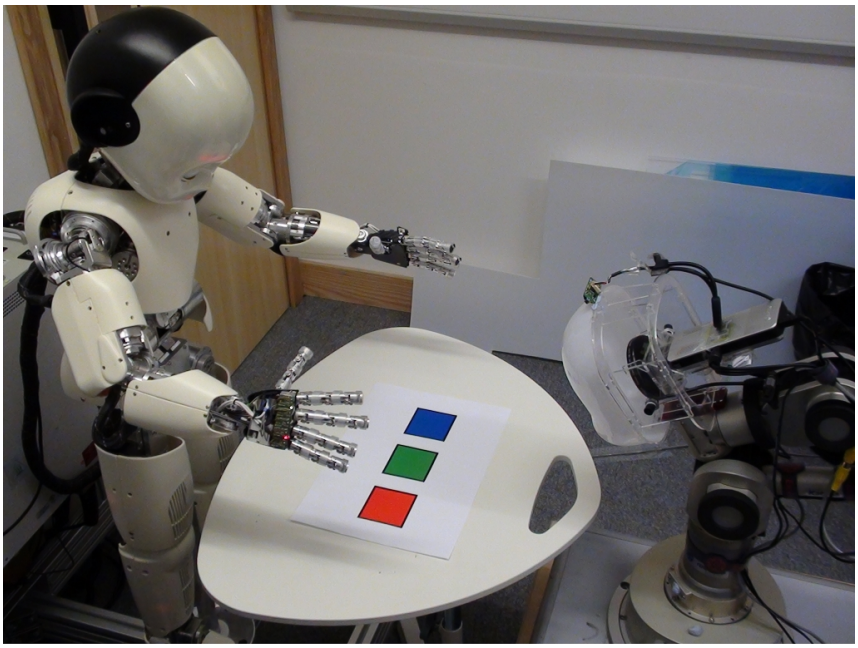


Figure 6.8: Robotic setup with the iCub robot on the left and the LightHead robot on the right examining a shared scene with colour stimuli.

6.2.2.2 Teacher-learner scenario

Experimental setup The set of stimuli as described above was used to generate sets of contexts consisting of 3 stimuli for language games. Two artificial agents representing the iCub and LightHead robot engaged in a series of language games. For each stimulus, the agent representing iCub would perceive the RGB values of a stimulus that was recorded from the iCub robot and in similar fashion the agent representing LightHead perceived the RGB values corresponding to the LightHead robot. The performance of the agents with robotic perception was then compared to the performance of agents with unaltered, normal perception. That is, agents in the normal condition perceived the RGB values from the items in the context in an unmodified manner, in the same vein as the agents with normal perception as described in section 6.2.1. Also in other aspects the experimental setup was similar to the setup as described in section 6.2.1.1.

Results Figure 6.9 shows the performance of the agents with robotic perception compared to agents with normal perception (two-sample t-test with $t(198) = -71.2809$, $p < 0.0001$). As can be observed, in the case of the real world stimuli, the agents with robotic perception perform still quite good compared to those with normal perception. The difference in perception causes a difference in performance but the agents are able to overcome these differences, resulting in an average communicative success of around 75% at the end of the language game interaction.

6.2.2.3 Population scenario

Experimental setup Next, an experiment was run in which a population of agents interacted through a series of language games to develop a common language system. All agents started with empty conceptual repertoires. The population consisted of 20 agents that were randomly assigned to represent either the iCub or LightHead robot (which determined the way they perceived stimuli). The stimuli were the same as used in the teacher-learner scenario. A series of language games ran for 50,000 cycles and 100 replications with these settings were run.

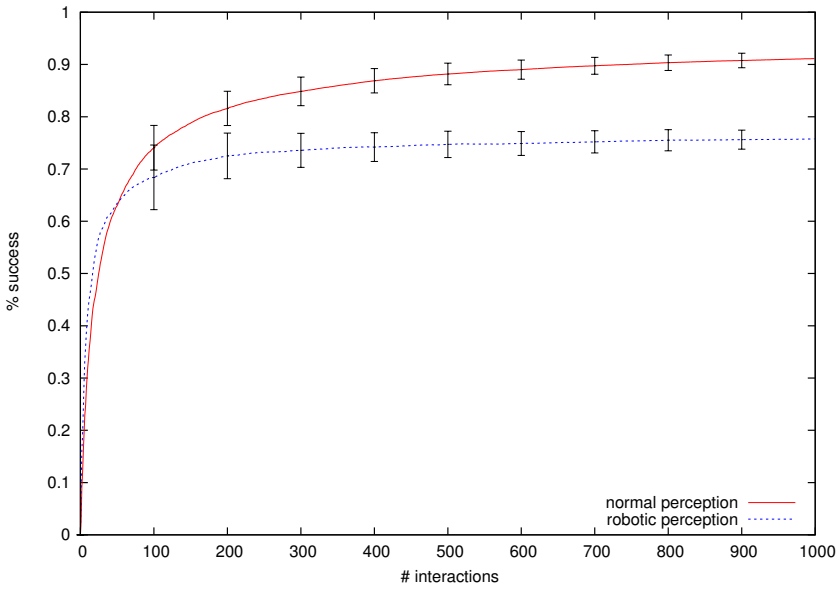


Figure 6.9: Performance of agents with robotic perception compared to agents with normal perception (error bars show SD).

Results Results indicate that compared to a setup in which agents perceive normally, the agents perceiving robot-recorded stimuli performed marginally better. A t-test shows a significant difference (two-sample t-test with $t(198) = 2.5389$, $p = 0.0119$), but communicative success for both cases is almost on par (figure 6.10). The reason that agents with individual perception perform this well is most likely the fact that there is no predefined knowledge in the system. Thus, agents are free to utilise word labels that are most effective for their environment and personalised perception.

Analysis The fact that performance is virtually identical to agents with normal perception is in contrast with the performance of the agents as described in section 6.2.1.2, as in the latter there is a difference between agents with individual perception compared to normally perceiving agents. This discrepancy is due to the fact that in the population scenario with robotic samples, effectively there are only two types of agents: one with iCub perception and another one with LightHead perception. In contrast, in the population as described in section 6.2.1.2, all 20 agents have individual perception, making it harder for the population to derive at a shared meaning system.

To verify this explanation, a simulation was run with the same settings as in

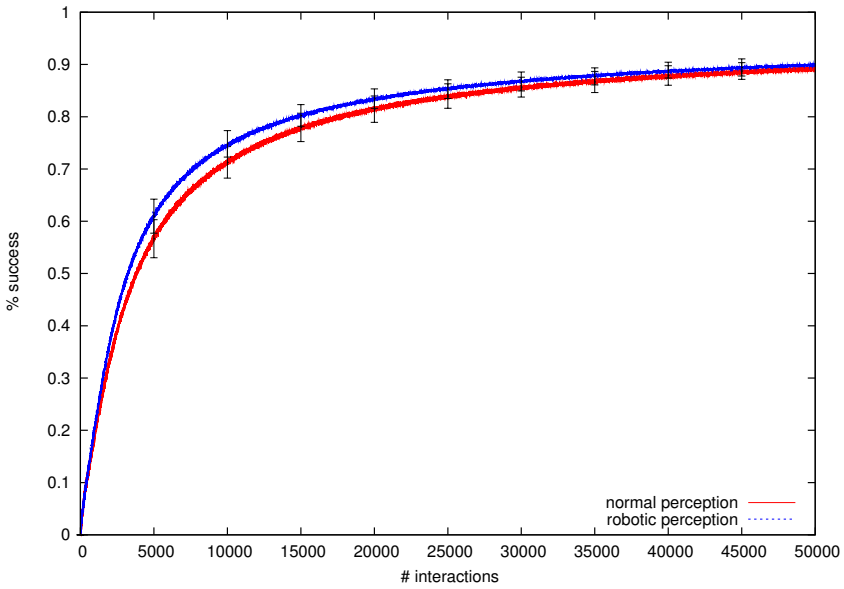


Figure 6.10: Performance of a population of agents with robotic perception compared to agents with normal perception (error bars show SD).

section 6.2.1.2, but with only two types of agents[1]. The results indicate that with only two types of individual perception, agents in the population perform as good as a population with normal perception.

6.2.3 Discussion

We have presented computational simulations in which agents with different perceptual capabilities successfully manage to develop a shared system of meanings. Perceptual differences in agents are modelled on physiological differences in the human vision system and on recordings from two different robots. Agents are able to overcome these differences in perception in both cases, in both a teacher-learner scenario and within a population of multiple agents.

The difference in perception has an influence on the effectiveness of communication, but through reshaping of colour categories agents are able to reach an acceptable level of communication. How does this come about? The answer to this question can be explained by considering the language games dynamics. In a language game, communication between two agents succeeds when the listening agent

[1]That is, the number of agents was still 20, but they were randomly assigned one out of two possible perceptual systems which were based on the difference in perception between the two robots.

is able to discriminate from the context the topic that is intended by the speaking agent. Crucial to this is the ability to adequately discriminate all items in the context, which in turn depends on how different these items are with respect to each other. Interesting cases arise when the listening agent is not able to discriminate the topic from the context that was intended by the speaking agent. Indeed, this is a vital part of the learning process, as it allows the listening agent to adjust its categories based on feedback generated through the failed language game.

When the perceptual capabilities of agents differ, it is to be expected that this situation will happen more often than when agents have identical perception. This is particularly the case, because when training data is generated, the notion of Minimum-distance-objects (D_{min}, section 3.2.7) governs the distance between all items in the context. However, when agent's perceptual capabilities are modified according to equation (6.2), it can easily be the case that for a context in which D_{min} is adhered to given normal perception, with the agent's particular manner of perception this is no longer the case. As illustrated in figure 6.3, the mapping from RGB to LMS can alter quite drastically the positions of objects in the colour space. As the manner in which agents' perception is modified contains a random set op weights (which is persistent for the agent), this effect may be more or less pronounced for different agents. Given the above, this will generate situations in which the label provided by the speaking agent does allow the learning agent to discriminate the topic from the context. This results, on average, in an increased number of failed communication events. This effect is more pronounced in the teacher-learner scenario, because in this the teacher has a fixed repertoire of colour concepts and word labels, and as such does not alter its word-concept associations, nor the coordinates of the colour concepts. In contrast, in the population scenario, all agents act both as teacher and learner. Thus, this leaves more freedom within the population to shape the agents concepts and word-concepts associations in response to the environment and the particular manner in which the agents perceptual capabilities are modified. Hence, a population is able to find a functioning system of meanings provided there

is some similarity in agents' perception[2].

This agrees with experiments in which human subjects were asked to play a language games using colours: agreement between subjects was on average 84% (Belpaeme, 2002a). We suggest that this mechanism of linguistic coordination may be analogous to how humans solve this problem, the latter also being a situation in which the link between perceptions and the words used to describe them is constantly reshaped through interaction with others. Furthermore, when the model is extended to robotic hardware, the same principles seem to hold.

Language games or, more generally, the adaptive coordination of categories through feedback, has been shown to overcome the diverging categories that would be caused by varying embodiment. When categories and concepts need to be coordinated between robots, language can act as the conduit for this coordination. The interaction between humans and robots also required coordinated categories and concepts, and although not demonstrated in the experiments discussed in this chapter, we propose the same process can also be used to coordinate the conceptual representation between humans and robots, which have perhaps the most extreme variations in embodiment.

6.3 Experiment 2

In a second series of experiments we explore the same issue as in experiment 1, that is, how agents with fundamentally different perception can achieve a common understanding of colour names. However, in this series of experiments, the manner in which agents perceive colours is more closely modelled on the human colour perception system; this is explained in more detail in section 6.3.1.

As such, while obtaining similar results, we argue that these experiments make a stronger case for the proposed explanation of how agents overcome perceptual differences through linguistic means. Agents' perceptual capabilities are based on human colour perception as described in section 6.1.1; as usual, their interaction

[2]When agents' perception is modified in a total random manner (but consistent per agent) a population performs little above chance.

type	c	μ	σ	R-square
S cone	0.9889	447.2	33.4	0.9928
M cone	0.9989	545.2	52.69	0.9965
L cone	1.0	567.9	64.78	0.9951

Table 6.1: Coefficients used in order to model responsiveness of S, M, and L cones.

with other agents is modelled through language games (section 3.2).

6.3.1 Modelling agents' colour perception

Artificial agents perceive external colour stimuli and communicate about this with other agents in a population. The agents colour vision proceeds as follows: an agent perceives a colour stimulus as a wavelength with a certain value w in nm. This value is fed into a function which was fitted as a Gaussian distribution on the cone responsiveness data as reported in Stockman and Sharpe (2000) so that activation value a_t with $_t \in \{S, M, L\}$ reflects the sensitivity curves for the three cone types S, M and L.

$$a_t = c_t e^{-\left(\frac{w-\mu_t}{\sigma_t}\right)^2} \tag{6.3}$$

Table 6.1 displays the coefficients used to model the activation levels of S, M and L cones, along with the goodness of fit (R-square). For each perceived w a triplet (a_S, a_M, a_L) is returned; figure 6.11 shows a projection of the S, M and L responses for different wavelength values in a LMS-responsiveness space.

The next step in colour vision is generally thought to be a projection of cone activations into two dimensions, the so called 'cone-opponent axes' specifying red-green and blue-yellow (Webster et al., 2000). To model this, the activation triplet (a_S, a_M, a_L) resulting from equation (6.3) is processed in the following manner. First, each activation is divided by the total level of activation:

$$(a_S, a_M, a_L) = \left(\frac{a_S}{a_T}, \frac{a_M}{a_T}, \frac{a_L}{a_T}\right) \tag{6.4}$$

with $a_T = a_S + a_M + a_L$. Then the opponency channels are created according to equation (6.5):

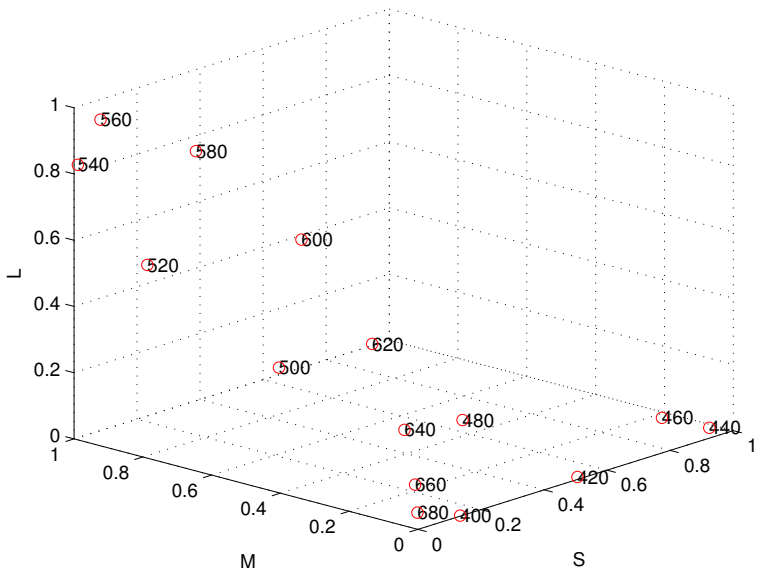


Figure 6.11: Response to various wavelengths projected in LMS space.

$$[x, y] = [(a_L - a_M), (a_S - (p_L a_L + p_M a_M))] \tag{6.5}$$

thus resulting in an $[x, y]$ coordinate set on the red-green by blue-yellow plane, where p_L and p_M are parameters that govern the relative input of the M and L cones. The value of p_L and p_M has been established to be about 0.5 (Sankeralli and Mullen, 1996), and so the blue-yellow axis can be thought of the activation of S vs M+L cones. Figure 6.12 shows a series of wavelengths projected on this cone opponency plane[3].

6.3.2 Language games applied to colour learning

Discrimination game - single agent Before agents can interact with other agents to establish a system of shared meanings, they first need to be able to classify incoming colour stimuli. This is based on the perceptual mechanisms as described in section 6.3.1 and combined with a discrimination game.

A discrimination game typically results in a division of the input space into a limited number of categories. The actual number of categories an agents ends up

[3]The background colouring is shown to give an idea of the various wavelengths and the corresponding colour. The actual colours would not be exactly as shown.

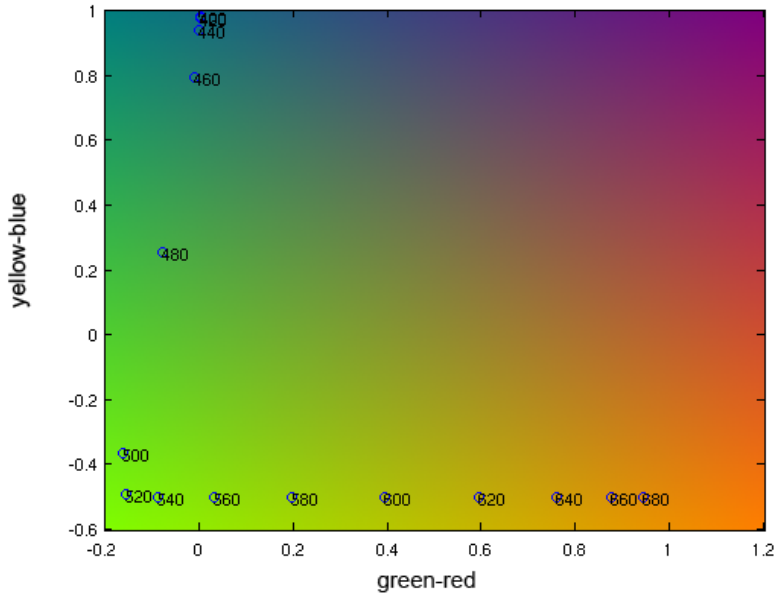


Figure 6.12: Response to an incremental series of wavelengths projected on the cone opponency plane.

with varies, as it is guided by a random selection of stimuli and governed by a parameter D_{min} which specifies the minimum distance between all stimuli in the context (see section 3.2.7). The smaller D_{min} is, the more categories an agent will end up with. Figure 6.13 displays an agent's representation of the input space after 1000 discrimination games with $N_{context} = 3$ and $D_{min} = 40$ projected in a LMS space (top) and against the visible spectrum (400-700nm, bottom).

Language game - population After agents have learned to categorise the input space, they can engage into linguistic interaction with other agents in the population. A baseline example of the results of such a language game is shown in figure 6.14. This illustrates the communicative success of a population of 10 agents ($N_A = 10$) after running 50,000 interactions which involved 2 randomly chosen agents. Thus, each agent participated on average in 10,000 interactions ($N_I = 10,000$). Other parameters were $N_{context} = 3$, $D_{min} = 40$ and $R = 25$. As can be seen in the figure, agents are able to achieve a good level of successful communication after sufficient interactions.

Another useful measure is the degree of similarity that exist between the agents' categories. To measure this we use the weighted sum of minimum distances (equa-

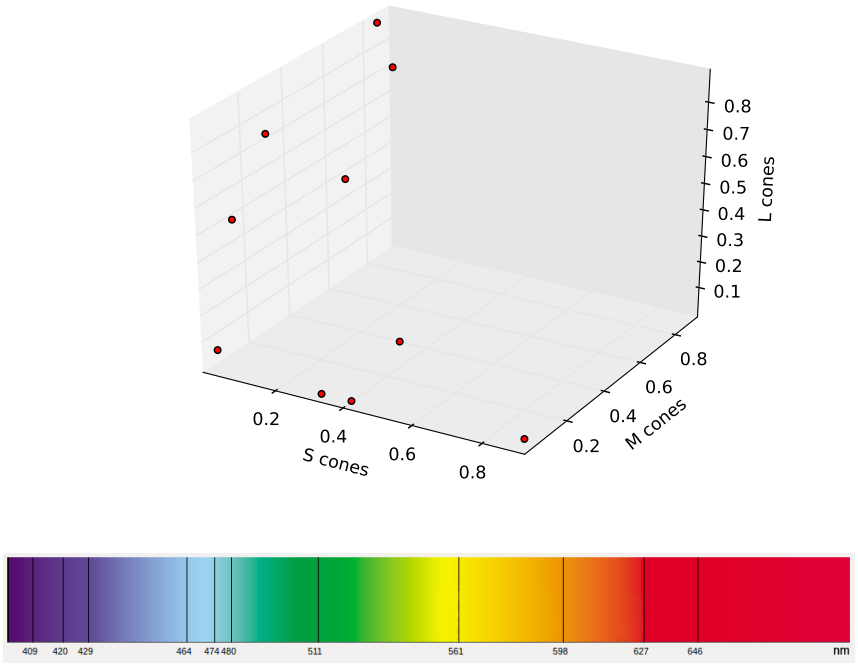


Figure 6.13: Internal categories of an agent after application of discrimination games projected in LMS space (top) and against the visible spectrum (bottom); lines indicate category boundaries.

tion (6.6); Belpaeme, 2002b) which calculates the distance $D_{set}(A, B)$ between two agents' sets of categories.

$$D_{set}(A,B) = \frac{\sum\limits_{a \in A} \min\limits_{b \in B} \|a-b\| + \sum\limits_{b \in B} \min\limits_{a \in A} \|a-b\|}{|A|.|B|} \tag{6.6}$$

$$D_{pop} = \sum_{a,b \in A} D_{set}(a,b) \tag{6.7}$$

with A and B being to sets of categories and the distance metric $\|a-b\|$ being Euclidean. This is then summed over the whole population (equation (6.7)), as to obtain an overall measure (D_{pop}) of how close all agents' categories are within a population. Application of this measure is shown in figure 6.15, which displays the distance between agents' categories which decreases over time, thus indicating that the agents' categories become more similar throughout language game interactions.

A more in-depth illustration of how agents' categories change over the course of participation in language games is shown in figures 6.16 and 6.17. Recall that agents' categories are shaped through the dynamic process of the language games

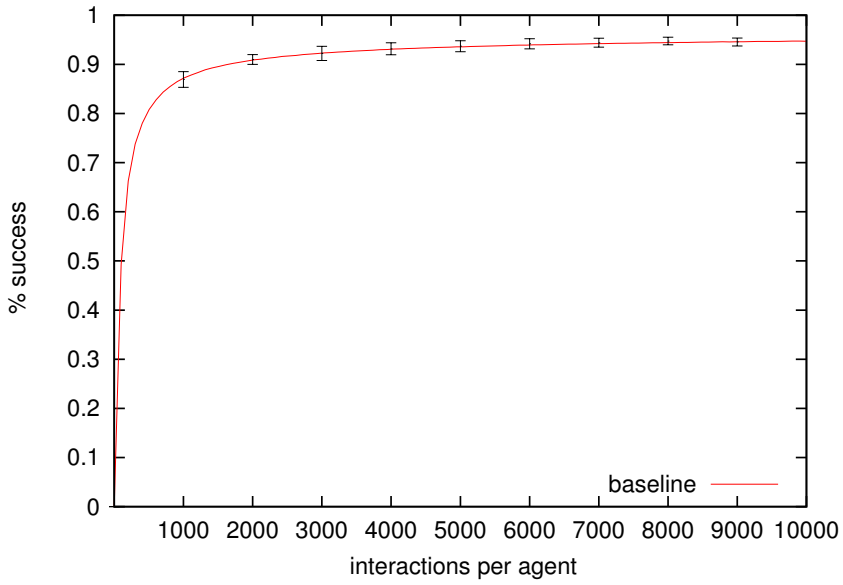


Figure 6.14: Communicative success of a population of 10 agents.

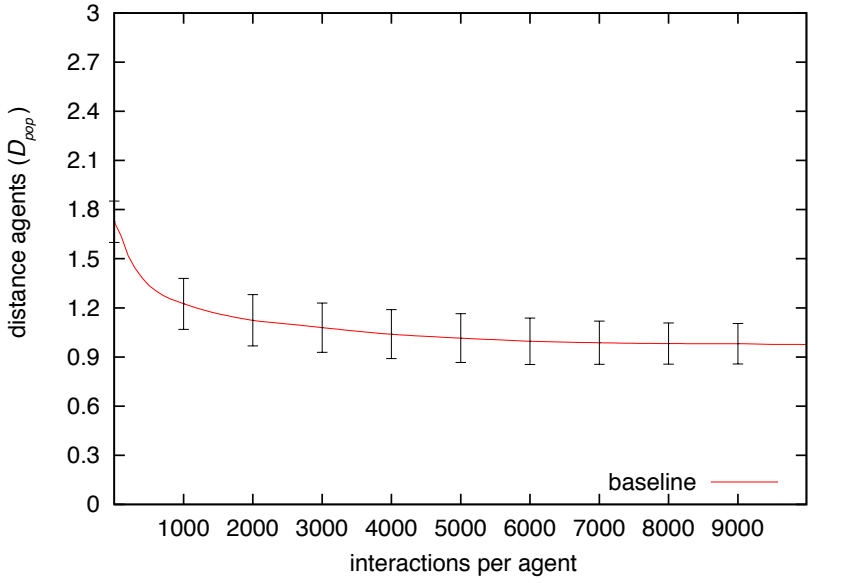


Figure 6.15: The distance between agents' categories D_{pop} in a population of 10 agents decreasing over the course of language games.

played within a population; depending on the outcome of multiple guessing games agents shift their categories and associated word labels, as governed by the dynamics described in section 3.2.4. Figure 6.16 shows the categories that the agents have acquired through discrimination games, which is essentially their subjective segmentation of the input space. To generate the figure, each agent is queried for the series of wavelengths [400, 700] and from this category boundaries are derived. In figure 6.17 the agents' categories are displayed after their participation in language games, thus their categories are shaped through this linguistic interaction. As can be observed from these figures, agents' categories resemble one another much more after linguistic interaction through language games.

6.3.3 Learning of colours with perceptual differences

To model differences in the way artificial agents perceive colour, their perception is modified in the following manner. Individual agents are endowed with a parameter $cone_r$ which specified the ratio of the three cone types. Thus, an agent with $cone_r = [0.1, 0.45, 0.45]$ has a cone ratio of 10%, 45% and 45% for S, M and L cones, respectively. When an agent observes a colour stimulus, a_t resulting from equation (6.3) is multiplied by the respective cone ratio $p_t \in cone_r$, as illustrated in equation (6.8).

$$a_t = p_t c_t e^{-\left(\frac{w-\mu_t}{\sigma_t}\right)^2} \tag{6.8}$$

This effectively results in a lower activation a_t, depending on the cone type t and the associated ratio. This assumes there is an additive mechanism at play that regulates the activation and subsequent processing of colour data in human colour perception. When all agents in the population have the same cone ratios, there is no effect on their ability to communicate effectively. If, for instance, all agents have $cone_r = [0.1, 0.45, 0.45]$, this effectively shrinks the S dimension to 10% and the M and L dimension to 45%, but it does not entail any functional differences as the range of the dimensions is arbitrary and the change is consistent for all agents.

However, if, like in humans, agents differ in their cone ratios on an individual

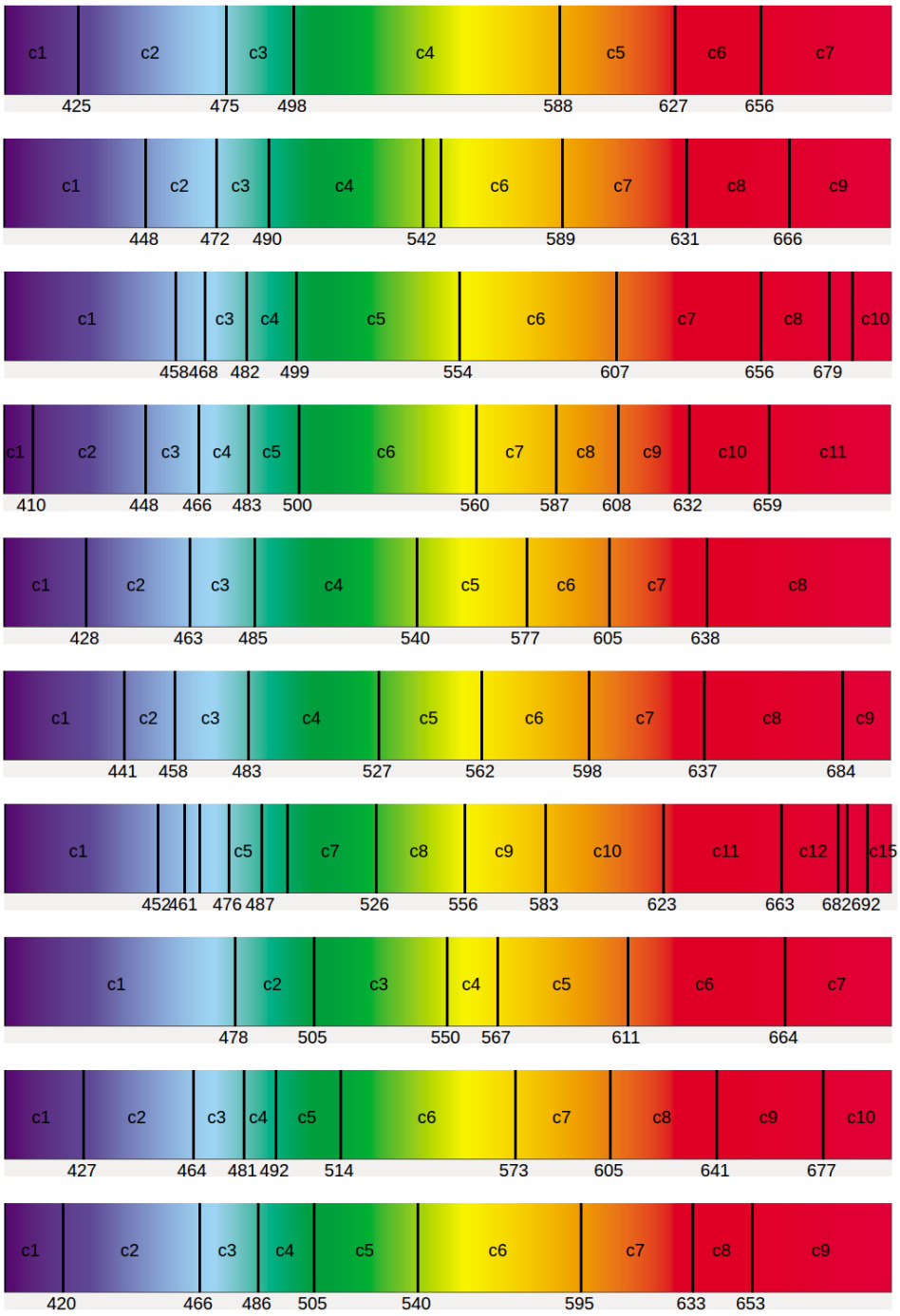


Figure 6.16: Categories of agent 1 to 10 (top to bottom) after playing discrimination games. Cx indicates category numbers; the wavelength is printed underneath the spectrum (when possible).

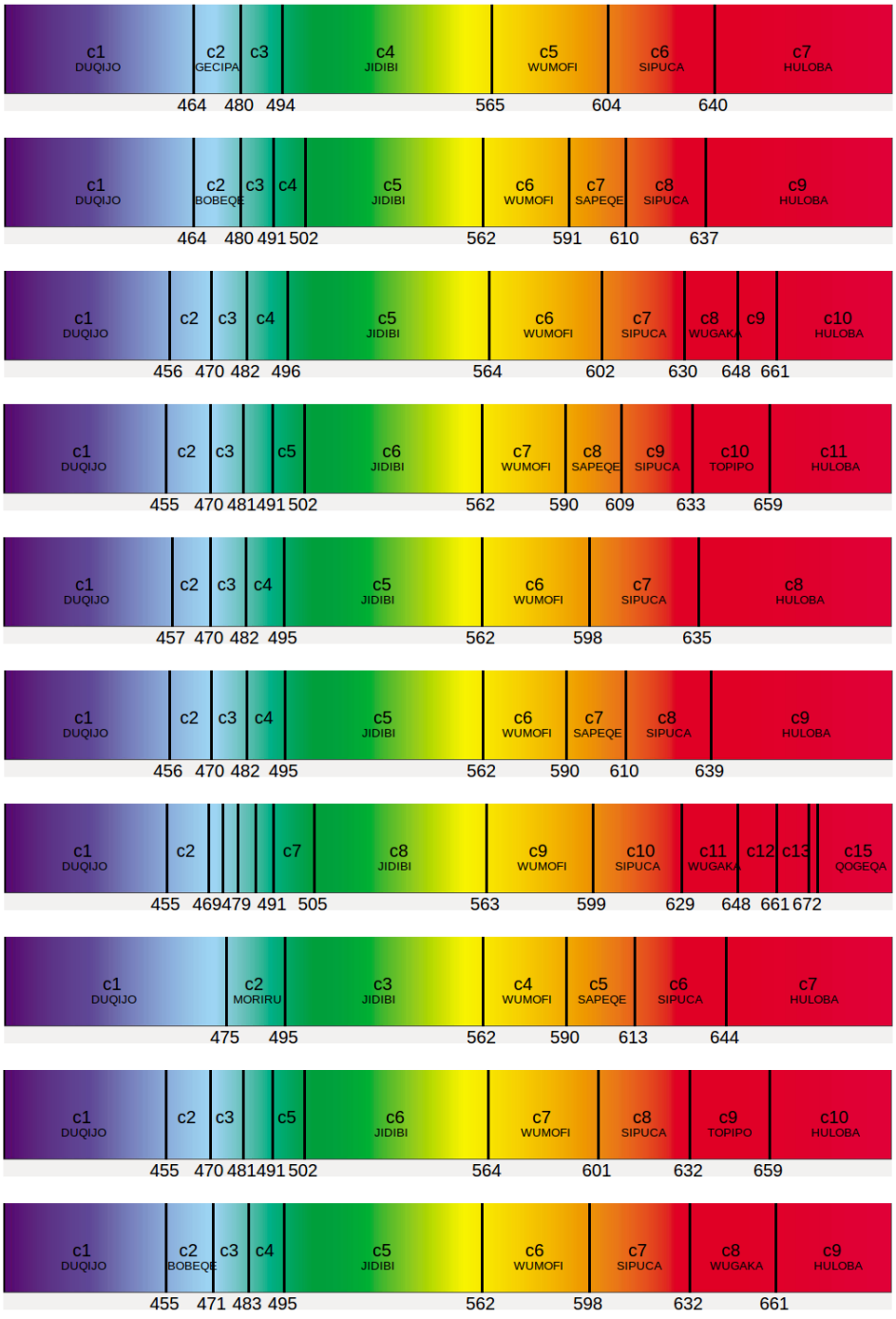


Figure 6.17: Categories of agent 1 to 10 (top to bottom) after playing language games. Cx indicates category numbers and the word label that is used within the population is displayed under this. The wavelength is printed underneath the spectrum (when possible).

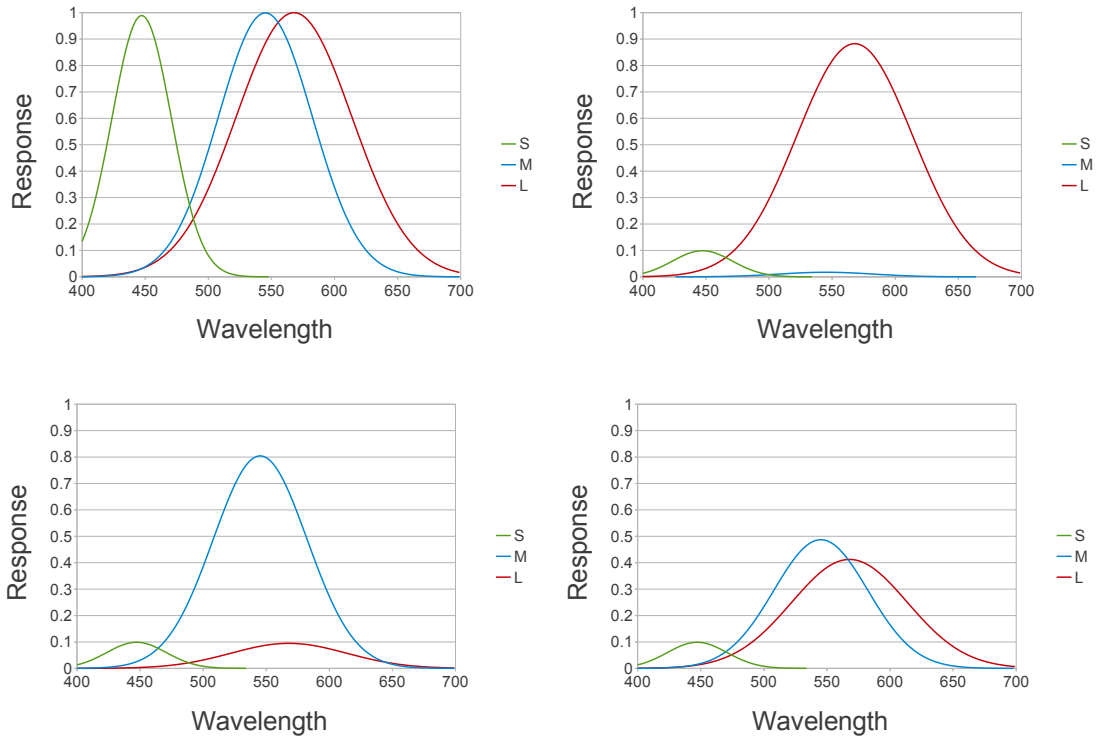


Figure 6.18: Cone response levels for a normal agent (top left) and three agents with random cone ratios (top right: $cone_r = [0.1, 0.02, 0.88]$, bottom left: $cone_r = [0.1, 0.81, 0.09]$ and bottom right: $cone_r = [0.1, 0.49, 0.41]$).

basis, this might have an effect of their ability to communicate. To test this situation, upon initialisation, the agents' $cone_r$ is randomly chosen in such a way that p_S is always 10%, but the remaining 90% is be randomly divided over p_M and p_L. This results in different cone responses for individual agents. The effect of this treatment is shown in figure 6.18 which shows the cone response curves for a normal agent, and agents with $cone_r = [0.1, 0.02, 0.88]$, $cone_r = [0.1, 0.81, 0.09]$ and $cone_r = [0.1, 0.49, 0.41]$ respectively.

To illustrate the effect of varying cone ratios on the categories that agents form during discrimination games, all agents categories are plotted in the cone opponency plane. As can be seen in figure 6.19, in the baseline condition the agents' categories are neatly following the same line as displayed in figure 6.12. However, as can be observed in figure 6.20, for the agents with random cone ratios the categories are much more distributed over the whole of the plane. The same effect is illustrated in figure 6.21. In here the categories of agents are plotted for the random cone ratio condition, after application of language games, for one replication only. What can

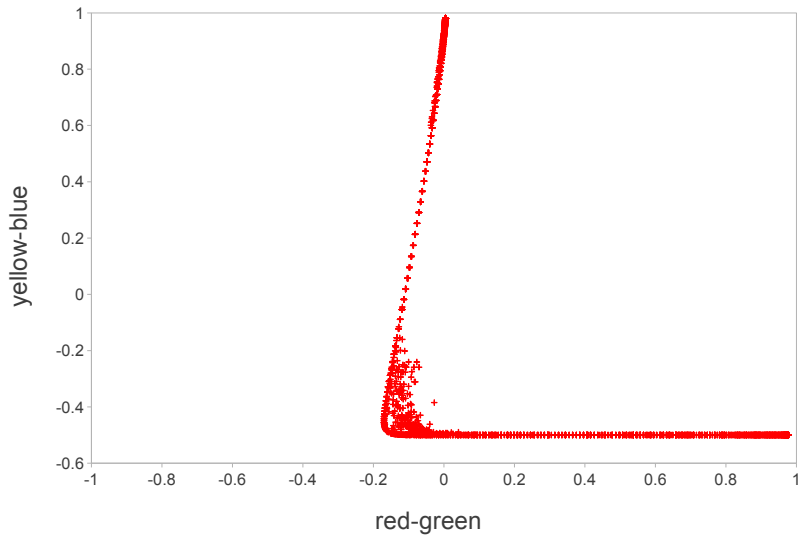


Figure 6.19: Categories of agents with equal cone ratios projected in the cone opponency plane. The categories are aggregated over all agents in the population for all 25 replicas.

clearly be seen is that each agents categories follow a different line through the cone opponency plane, depending on the cone ratio of the particular agent.

The colour yellow is said to be experienced in the absence of red and green, so this occurs when the activation of M and L cones cancel each other out. Based on the colour perception models as described above, for agents that have equal cone ratios this is the case for a stimulus of 555nm[4]. For agents with random cone ratios the stimulus at which activation of M and L cones cancel each other out varies widely for each individual agent, as the agent's individual cone ratio dictate the activation strength. To illustrate this point, all agents in the population were probed with a range of wavelength stimuli and their cone response levels (S, M and L) were observed. Then, to identify the stimuli for which resulting activation of M and L cones cancel each other out, the following rule was applied:

[4]While 555nm is the wavelength at which M and L cones cancel each other out for the particular colour models that are used here, in general a wavelength of about 580nm is considered yellow by most people. This is presumably the case because additional weighting mechanisms play a role in the processing of colour perception such as discussed by (Neitz et al., 2002, p.787). However, equation (6.5) is used to calculate coordinates in the cone opponency plane for a given wavelength, which assumes no such weighting mechanism. As such, a stimulus of 555nm results in M and L cancelling each other out. However, this poses no problem for the point that is made here, which is to show that for agents with varying cone proportions the wavelength at which M and L cancel each other out varies widely across the visual spectrum.

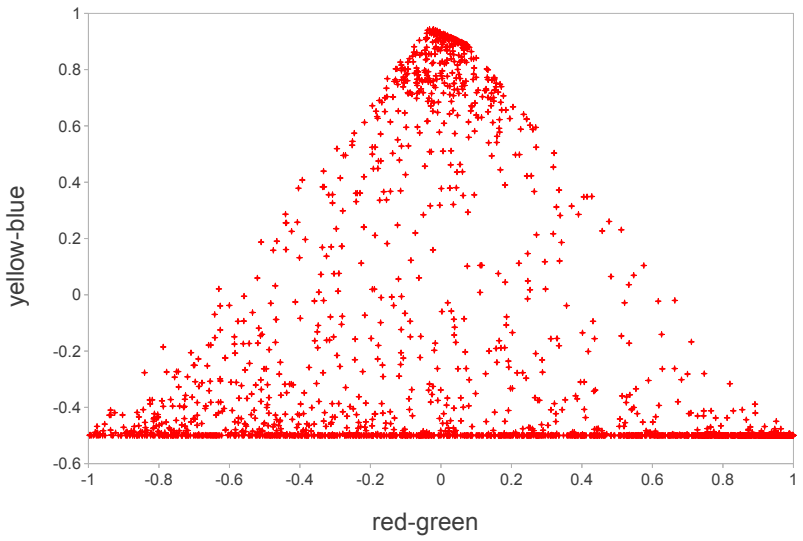


Figure 6.20: Categories of agents with random cone ratios projected in the cone opponency plane. The categories are aggregated over all agents in the population for all 25 replicas.

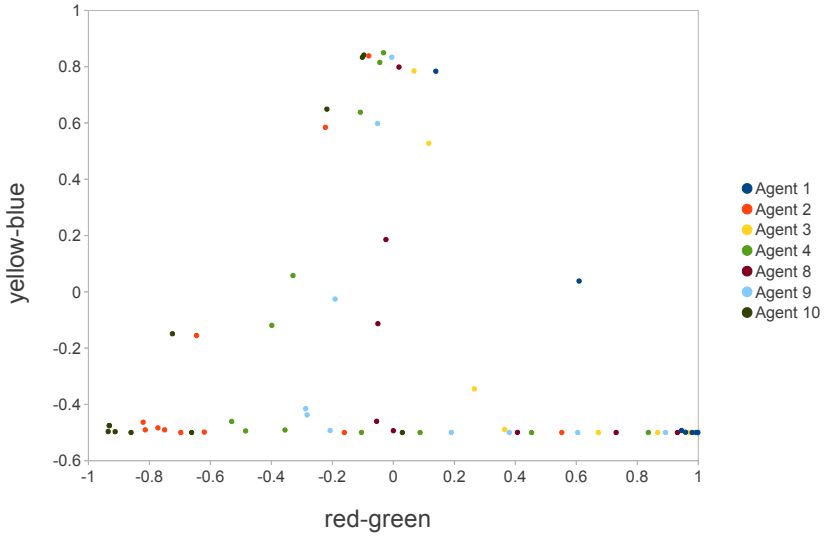


Figure 6.21: Agents categories projected in the cone opponency plane for agents with random cone ratios after language games. To keep the figure readable, not all agents from the population their categories are plotted.

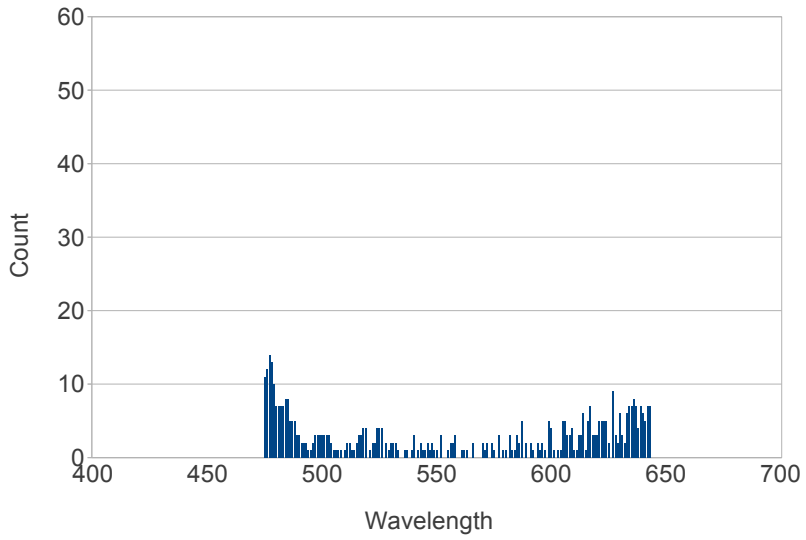


Figure 6.22: Display of stimuli (in wavelength) for which L and M cone activations cancel each other out (L-M=0) for agents with random cone ratios. The figure shows a cumulative plot of aggregated stimuli that fit the rule as formulated in equation (6.9) for all agents in the population, for 100 replicas.

$$(S < 0.05) \wedge ((M + L) > 0.05) \wedge (|L - M| < 0.001) \tag{6.9}$$

In words, this rule selects those stimuli that result in an agent's cone activation levels such that S is not active (< 0.05), M and L are active (> 0.05) but the absolute difference between M and L is very small (< 0.001). The stimuli for which this was the case were aggregated over all agents in the population for 100 replicas, and cumulatively plotted, as shown in figure 6.22. What can clearly be seen is that due to the varying cone ratios of the individual agents, the stimulus (in wavelength) at which L and M cone activations cancel each other out varies widely amongst different agents.

6.3.4 Effects of perceptual differences

The results of learning through language games by agents with varying cone ratios are shown in figure 6.23 and figure 6.25 (left), which compares the resulting communicative success to the baseline performance. Parameters for this simulation were the same as the simulation described in section 6.3.2. Figure 6.25 (right) shows a coherence test, in which the word labels that are used within the populations are

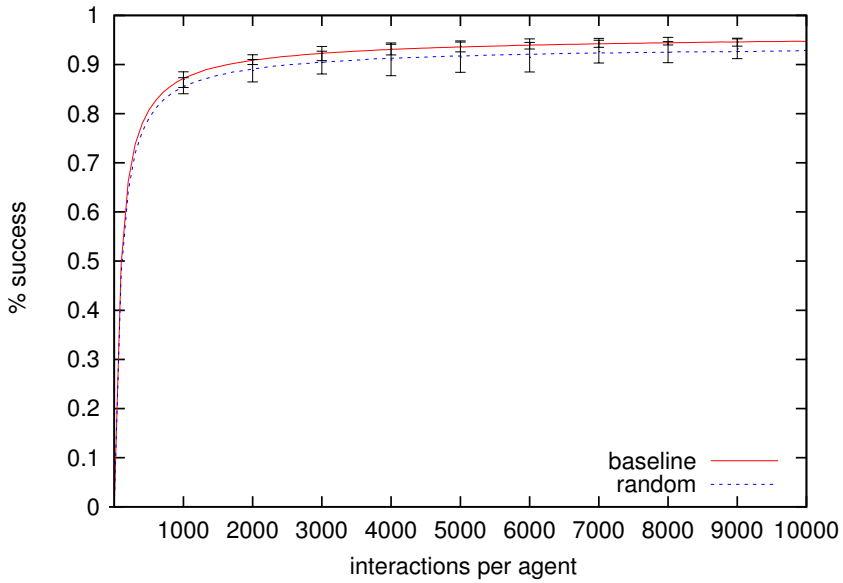


Figure 6.23: Communicative success of a population of agents with random cone ratios compared to the baseline performance.

analysed[5]. Agents are also tested on the level in which they agree on words used for randomly picked stimuli. For 1000 tests 2 randomly chosen agents both stated their word label for a randomly drawn stimulus. If the words are the same, they score a point. This provides a measure of word agreement. The results over all replicas after application of language games are 0.74876 for the baseline condition and 0.74624 for the random cone ratios condition. This is not significantly different (two-sample t-test with $t(48) = 0.1848$, $p = 0.8542$).

What can be observed is that the communicative success of agents with varying cone ratios is very close to the communicative success of agents with equal cone ratios. However, when looking at the distance between agents' categories (figure 6.24), it is striking that the agents categories are much more different compared to a population with equal cone ratios. Moreover, whereas in the latter case the population distance gradually decreases, this is not the case for agents with differing cone ratios. This reflects the fact that these agents have a different segmentation of the underlying cone opponency plane.

Next, we determined how similar the perceptual categories are for what agents

[5]The coherence test works as follows. For each stimulus in nm all 10 agents in a population list the word label that best matches the stimulus. The coherence score C is 1 when all agents in the population use the same word, and 10 when all agents use a different word. Coherence is then calculated as $1 - ((C - 1)/9)$, so 1 for maximum coherence and 0 for no coherence. This is then averaged over all replications.

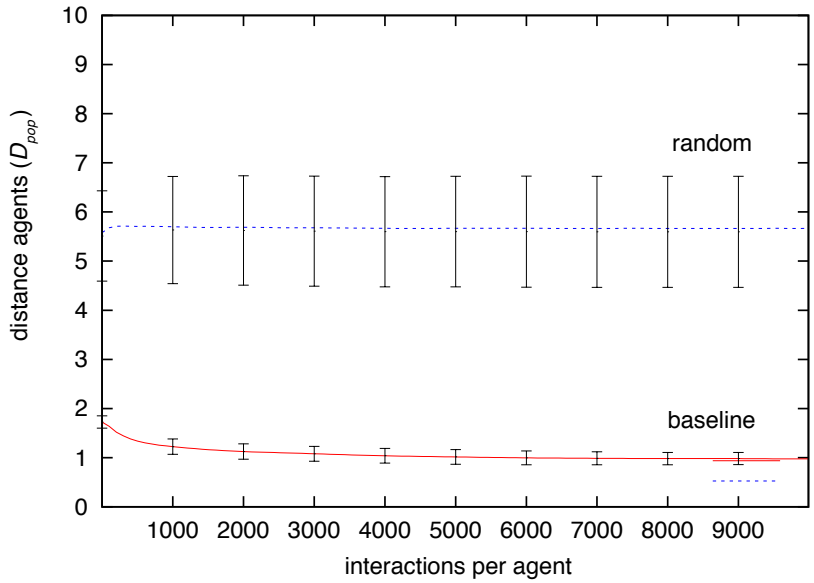


Figure 6.24: Distance between categories of a population of agents with random cone ratios compared to the baseline performance.

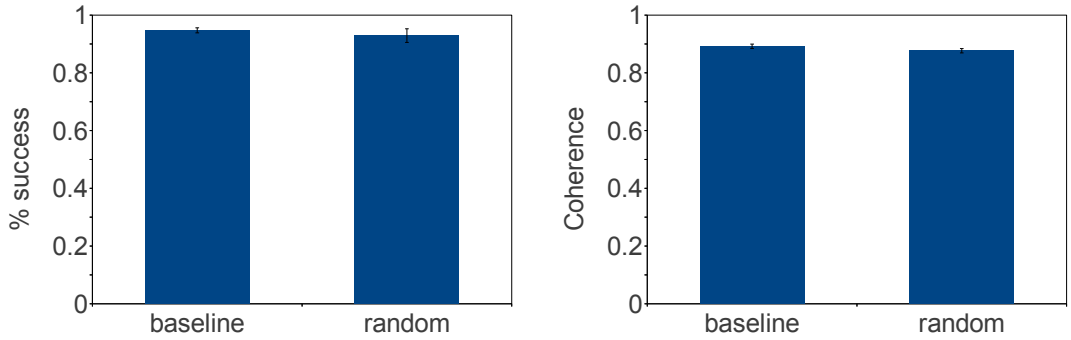


Figure 6.25: Comparison of population of agents with random cone ratios to the baseline performance (left), differences are small but significant (two-sample t-test with $t(48) = 3.6139$, $p = 0.0007$), and within population coherence for all replicas (right), for which the difference is also significant (two-sample t-test with $t(48) = 7.2507$, $p < 0.0001$).

perceive as 'yellow'. This is done by recording which perceptual category responds (i.e. is closest) to a stimulus of 580nm[6]. The coordinates of this matching category are plotted in the cone opponency plane. This is done for all agents in a group for all replications. There are four groups: agents with normal cone ratios before playing language games (baseline before), after language games (baseline after), agents with random cone ratios before language games (random before) and after language games (random after). The aggregated categories are plotted in figure 6.26. What can be

[6]Generally a wavelength of about 580nm is considered 'yellow', but from an agent's perspective querying a particular wavelength is arbitrary.

observed is that in the baseline case the application of language games causes the categories associated with the 580nm stimulus to become a bit more similar with respect to other agents in the population. However, this is not the case in the random condition; in here after language games the categories associated with 580nm are still scattered throughout the full range of [-1,1] in the x-axis of the cone opponency plane. Thus, in this case the agents maintain their individualised representation of the category and only align with other agents on the linguistic level.

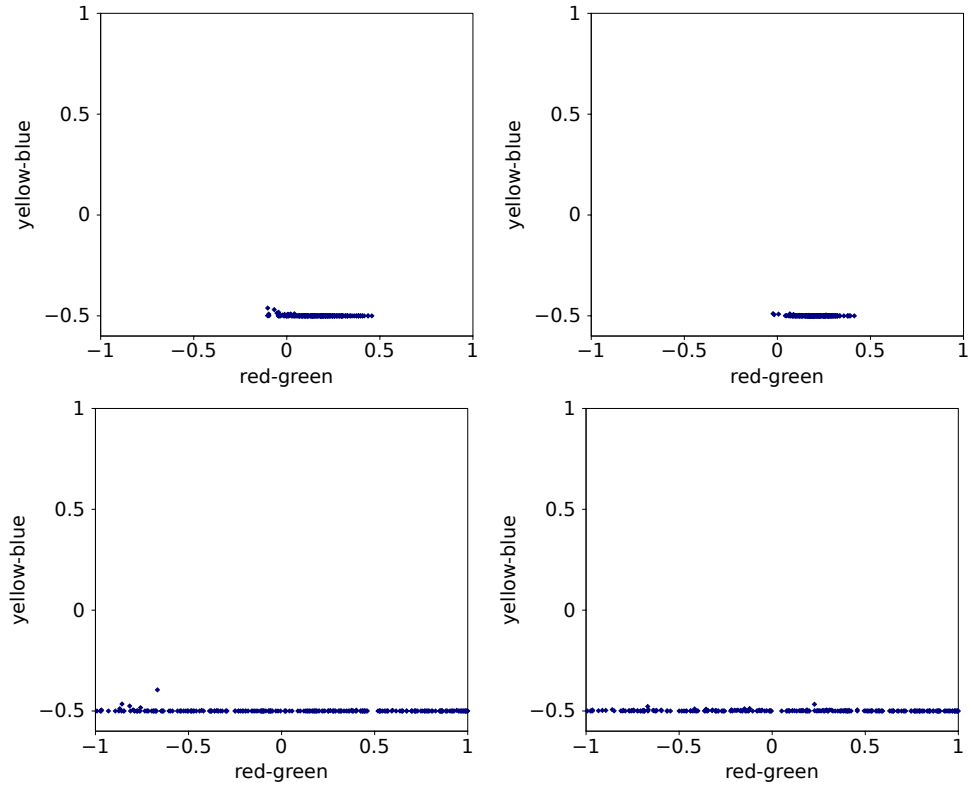


Figure 6.26: Perceptual categories responding to a stimulus of 580nm plotted in the cone opponency plane. Top displays baseline before (left) and after (right) language games, after which the categories have become slightly more focussed; bottom displays random cone proportions before (left) and after (right) language games.

The alignment on the linguistic level (after the agents have engaged in language games) is illustrated in figure 6.27. In here agents with random cone ratios are probed with a stimulus of 580nm to represent 'yellow'. For the word that is most strongly associated with this stimulus (the referent) the spread is measured (i.e. the range of stimuli in wavelength with which this word is associated). The centre of this spread is cumulatively shown in the figure, for all agents in a population for all replicas. As can be observed, the word label that is associated with the stimulus

representing ‘yellow’ is very much focussed in a small range of wavelength. Thus, compared to figure 6.22, the agents have focussed their perception of ‘yellow’ on the linguistic level to a much smaller range.

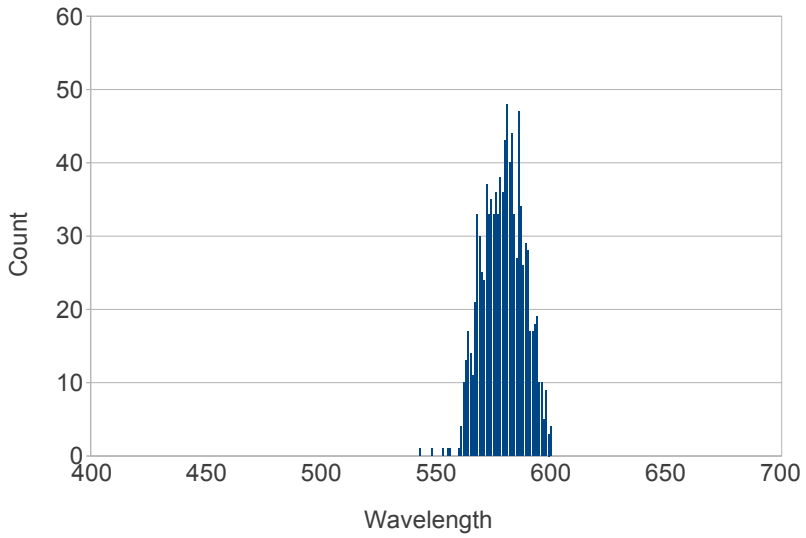


Figure 6.27: Cumulative display of centre of spreads (in wavelength) associated with the words that respond to a 580nm stimulus for agents with random cone ratios. The figure displays the aggregated count for all agents in a population for all replicas.

The application of language games for a population in which agents have variations in their cone ratios results in a group of agents which have quite dissimilar categories, but are yet able to achieve a level of communicative success that is very close to a population of agents with equal cone ratios. The ability to reshape the meaning of word labels used within the population proves robust against dissimilar underlying perceptual categories. Moreover, whereas before linguistic interaction agents perception of ‘yellow’ varies widely (figure 6.22), after engaging in language games the range in nm which is associated with ‘yellow’ is much more focussed within the population (figure 6.27)

6.4 Chapter summary

Given the embodiment thesis, one would expect that differences in physical embodiment would impact on cognition and perception. Despite physiological differences in the human retina, in particular the ratio of M and L cones, people behave very

similar in colour naming tasks. Strikingly, people with very different M:L cone ratios show only a small variance in the range of wavelength that they perceive as unique yellow. It has been suggested in the literature that a kind of neural weighting is responsible for this alignment of colour terms, but how this exactly works remains unclear.

In this chapter we have suggested an alternative account in which it is not lower level neural weighing, but a higher level dynamically shaped linguistic system that enables people to align their colour experience. We have shown the feasibility of this account by demonstrating how agents with perceptual differences based on physiological differences in the human retina and data recorded from two robots are able to achieve an effective communication system. By dynamically reshaping the meaning of word labels used to describe colour stimuli, it is not necessary for agents to have an identical low level perceptual organisation. Indeed, individual agents can have significant differences in their perceptual properties and yet 'speak the same language'; that is, reach an effective level of communication. We argue that given the highly intersubjective nature of human experience, it is only natural to assume a constant reshaping of linguistic meanings based on interaction with others. The relatively simple experiments which capture these linguistic dynamics provide an account of how perceptual differences can be overcome for communicating agents.

Chapter 7

Social learning with robots

This chapter describes experiments with the computational model involving humans and robots. Rather than having simulated agents performing the interactive learning, the interlocutors are humans and (embodied) robots. Real world implementations of models can increase their strength because successful application shows the feasibility of actually working 'in the wild', as opposed to working in a controlled virtual environment. The real world is always much more noisy, chaotic and less structured than any simulated environment; real world applications are therefore more credible than their simulated counterparts, as they need to be more robust.

This chapter combines the findings from chapter 5 with the insights from human-robot interaction as described in section 1.5. Particularly the aspect of interactive social learning is explored. First, social learning within an HRI context is explored, focussing on crucial capabilities such as gazing and category learning. Then a description of a new kind of robot, LightHead, is provided. This robot is tailored towards HRI and sports a retro-projected animated face which has many advantages over more classic mechatronic robotic faces. The LightHead robot was used in an exploratory study which assessed people's abilities to read the robot's gaze. Having established the LightHead robot as a functional platform for HRI, experiments are then described in which a learner (both in simulation and embodied in the robot) employs active learning strategies while learning from human teachers.

7.1 Human-Robot Interaction

Traditionally, Human-Robot Interaction (HRI) has focussed on industrial applications in which people operate robots to perform various tasks. As such, robots are perceived as tools and human-robot interaction research focussed on interfaces, feedback, sensory information and control. However, due to a steady increase in robots' sophistication in recent years, application domains have expanded drastically and subsequently HRI has become more complex. HRI as an independent field of research started to emerge in the mid 1990s and early years of 2000 (Goodrich and Schultz, 2007). Rather than viewing robot as mere tools, Breazeal (2004) identified four different paradigms of HRI: 1) robot as tool, 2) robot as cyborg extension, 3) robot as avatar and 4) robot as sociable partner. Each of these paradigms entail a different view of the role of the robot within the interaction: in the first the robot is seen as a piece of (complex) equipment, in the second the robot can be seen as an extension of ones own body, in the third the robot is a surrogate for some other person and in the fourth paradigm the robot is seen as a social entity. In the last two paradigms, people interacting with a robot would not use specialised interfaces, but rather rely on the same interaction channels that are used when interacting with other people; e.g. verbal communication[1], mutual gaze understanding, pointing, gestures and facial expressions. Particularly the fourth paradigm, in which a robot is perceived as a social partner, opens up new areas of research and applications, by placing emphasis on the skills and attributes robotic systems would need to posses in order to interact socially with people (Dautenhahn, 2007b).

Also within this last paradigm, where robots are seen as social entities that can interact with people, different themes and objectives have emerged. For instance, some HRI research takes a developmental perspective (akin to the developmental robotics approach as described in section 1.3), in which robotic babies are used to investigate developmental aspects of cognitive systems and child/robot-caregiver relationships (Minato et al., 2007; Ishihara et al., 2011). Other research has fo-

[1]Verbal communication with an artificial system requires natural language processing (NLP), which constitutes a whole research field in itself. As of to date, no artificial system can cope with open ended natural speech; applications tend to be restricted to limited domains.

cussed more on human-robot interactions in general, in which the objectives are to investigate people's expectations and manners in which robots should be designed and behave in order for society to accept them. Examples of such themed work are e.g. the lifelike androids such as Ishiguro's Geminoid (Nishio et al., 2007) and Hanson Robotics' Albert HUBO (Oh et al., 2006), the work of Goetz et al. (2003), which shows how people's compliance with robots improves when the robot's appearances and behaviours match their capabilities, and Kanda et al. (2004), who studied how interaction between robots and children could lead to the formation of social bonds. Related to this, benchmarks have been proposed (Kahn et al., 2007) to establish a means of measuring the success of building more and more life-like robots. Anticipating more integrated roles for robots in society, Takayama et al. (2009) investigated aspects of human-robot conflict management. These developments in HRI have sparked discussions of ethical concerns related to an increase of robotic presence and related HRI in society, e.g. Sharkey and Sharkey (2010, 2012), although others have argued that it might be too early to worry about such issues (Belpaeme and Morse, 2010).

Another 'branch' of HRI that has been on the rise is the application of robots in teaching and therapy, such as the robot therapy for elderly with dementia using a baby seal robot (the Paro robot, Wada et al., 2005), therapeutic robots aimed at interacting with autistic children (Dautenhahn et al., 2009; Kim et al., 2012), or the development of robots that are capable of prolonged interaction with diabetic children (the ALIZ-E project, Belpaeme et al., 2013), providing companionship to the child, as well as teaching them important concepts about diabetics and self-management. In similar vein, robots have also been used to facilitate teaching of healthy children (Tanaka and Matsuzoe, 2012). Recently, the topic of long-term interaction has gained attention as well (Leite et al., 2013).

7.1.1 HRI topics

To support effective and satisfactory interaction between robots and people, the robots need to be equipped with the right kind of sensory apparatus and means to

process the various datastreams that govern the interaction; not only in technical terms (e.g. sufficiently high resolution of camera images), but also the appropriate programming and design architectures, i.e. the robot's 'AI'[2].

As HRI by its very nature emphasises aspects of interaction, it draws attention to the variety of functions and mechanisms that facilitate interaction between humans, and that, as such, are perceived as natural by people. Humans are a linguistic species, and considering the prominent role of language in cognition (section 1.4), one might expect that language would be a prime factor for human interactions. However, a lot of aspects that facilitate and influence interaction between people (Hinde, 1972; Ekman and Friesen, 1981) and, by extension, people and robots (Breazeal et al., 2005; Han et al., 2012), are non-verbal. Examples of such non-verbal interaction mechanisms are for instance facial expressions (Gonsior et al., 2011), gestures and pointing (Yang et al., 2007; Sato et al., 2007), joint attention (Nagai et al., 2006) and eye gazing (Yoshikawa et al., 2006).

The latter topic, eye gazing, will be the object of interest for some of the HRI experiments presented in this chapter. Gaze understanding is considered very important for effective communication, as it plays a crucial role in social attention (Langton et al., 2000). As such, it is a relevant topic for HRI. A lot of research has been focussed on how robots can read gaze from humans, which is effectively the extraction of a gaze angle from a video stream and thus fundamentally constitutes a computer vision problem. Various algorithms have been proposed, e.g. (Atienza and Zelinsky, 2002; Yoo and Chung, 2005), for an overview see (Morimoto and Mimica, 2005; Hansen and Ji, 2010). However, as interaction is bi-directional, it is equally important that a person who is interacting with a robot is able to interpret the robot's gaze; particularly for establishing joint attention. The issue of establishing bi-directional eye-contact has been addressed in some studies (Miyauchi et al., 2004, 2005), but what remains still unclear is what factors influence people's abilities to infer where another (artificial) agent is looking? What is the influence of the physiognomy of an agents face and eyes on the users ability to infer where it is looking? And

[2]Of course, the field of AI is entails a very wide scope of topics (see e.g. section 1.2 and section 1.3), but from the current HRI perspective one might understand AI as the robot's control mechanisms that allow the robot to interact with people.

do the dynamics of eye movements have an influence? To address these questions an experiment was conducted in which we tested participants' abilities to read the gaze from a (novel) robotic head. This experiment will be presented in section 7.3. The issue has subsequently been recognised by the HRI community, as, following our study, several others have conducted similar experiments (e.g. Al Moubayed et al., 2012; Onuki et al., 2013).

Another topic that has gained more attention within HRI recently is social learning through interaction. This has been discussed in section 1.5.1. Taking these ideas as inspiration, a social learning experiment was conducted that combines the various topics of this thesis by having a robot learn concepts interactively from people (section 7.4).

7.1.2 HRI methodologies

As HRI studies tend to involve human participants, it is common to use techniques and methodologies derived from sociology and psychology. For instance, in a typical HRI study groups of participants would interact with a robot exhibiting a certain kind of behaviour, and a control group would be interacting with a more 'neutral' robot without this particular behaviour. Alternatively, a group of human participants would interact with various kinds of robots in sequence. Studies can be both qualitative and quantitative; data gathering and measurements are, apart from the robot log files, similar to those used in psychology as well, e.g. annotation and coding of audio and video recordings, asserting participants' performances through tests, questionnaires, response time and physiological aspects such as skin conductivity. Kidd and Breazeal (2005) discusses several measurement types for HRI, along with some design recommendations. However, as argued by Dautenhahn (2007a), the field of HRI is relatively young and rapidly evolving. As such, no true established methodologies have been defined that can be agreed upon by all involved. Even though efforts have been made to establish common metrics for HRI studies (e.g. Steinfeld et al., 2006), these metrics would not be applicable to all experiments, as the field is constantly evolving. Another proposal for more unified evaluations meth-

ods within HRI is the theoretical and methodological evaluation framework USUS (Weiss et al., 2009), which "addresses usability, social acceptance, user experience, and societal impact of humanoid robots used in collaborative tasks...".

The use of questionnaires is very common in psychology, where it is often used to measure the users attitudes; as such a large body of literature on their use exists (e.g. Groves et al., 2004). Within the field of HRI, the use of questionnaires are common as well; see e.g. (Walters et al., 2005) for an extensive discussion of questionnaire design in HRI studies. Advantages are for instance that they are rapidly applicable, and allow for quantitative analysis of participants' demographics, attitudes and their subjective experience regarding interaction with a robot. A drawback is for instance that questionnaires are typically filled out by participants alone, and as such do not allow for more in-depth questioning about some interesting details the participant disclosed. Hence, a qualitative interview might be more appropriate in some occasions. Bartneck et al. (2009) provide an elaborate discussion regarding questionnaire use in HRI, including common pitfalls regarding design, and offer tools and guidelines for their appropriate use as a methodology in HRI, aiming to make results from HRI studies more comparable.

HRI experiments can be very challenging to set up, because of the advanced technologies that tend to be involved and the interaction with human participants. This makes it often quite hard to reproduce a certain finding, which can be problematic from a scientific viewpoint. Novel methodologies have been proposed that aim to facilitate design, prototyping and testing of HRI scenarios (e.g. Woods et al., 2006). Another technique that is rather specific to the field of HRI, is the so called *Wizard of Oz* (WoZ) setup. Because of technical limitations, it is fairly common in HRI studies to employ a WoZ setup, in which (part of) the robot behaviour is controlled by a human experimenter; typically in such a manner that the human subject with whom the robot is interacting is not aware of this. Although this contains an element of deceit, it allows to study human-robot interactions that would otherwise not be possible because of limited capabilities on the robot side (Riek, 2012).

7.2 The LightHead robot

The CONCEPT project[3] studied how robots could learn concepts from a human teacher in an interactive manner. Part of this project was the development of a new kind of robot face, one that is very much tailored to support some of the crucial aspects of HRI. That is, non-verbal communication through facial expressions, mutual gaze understanding and sophisticated and coherent timing of actions. The robot face, dubbed LightHead, was developed by Frédéric Delaunay as part of his PhD research. It offers a considerably cheaper alternative to the more traditional mechatronic robot faces due to the use of off-the-shelve components.

LightHead has the appearance of a young child (see figure 7.1). Its main feature is the utilisation of *retro-projected animated face* (RAF) technology (Delaunay et al., 2009): the use of a small scale projector to project a computer generated character inside a semi-transparent facial mask. As such, the robot offers advantages over the more classic mechatronic faces, most notably the ease of projecting computer animations which allow for flexible character design and rich social interaction. The animated face projected into the mask is generated in real-time by an off-board computer, which typically is also used to control the 6-DOF robotic arm on which the face is mounted. Examples of some different facial expressions are shown in figure 7.2. Being computer animated, the facial appearance and expressions of the robot are very versatile and can be modified on the fly, depending on the application and context in which the robot is used. Besides 'regular' facial expressions, display of more subtle communicative signals such as an animated tongue, iris dilatation, blushing and other socially salient cues can easily be achieved.

To support the interaction, the robot head is mounted on a robot arm such that the arm acts as a thorax and neck. This allows the robot's head to move, thus giving the impression of looking around or craning over a table, for example to examine objects in front of it. For more details regarding the design, materials, software architecture and implementation, see Delaunay (2014).

[3] `http://www.tech.plym.ac.uk/SoCCE/CONCEPT/`

Figure 7.1: The LightHead robot face, mounted on a robot arm (Jennie Hills, Science Museum, London).

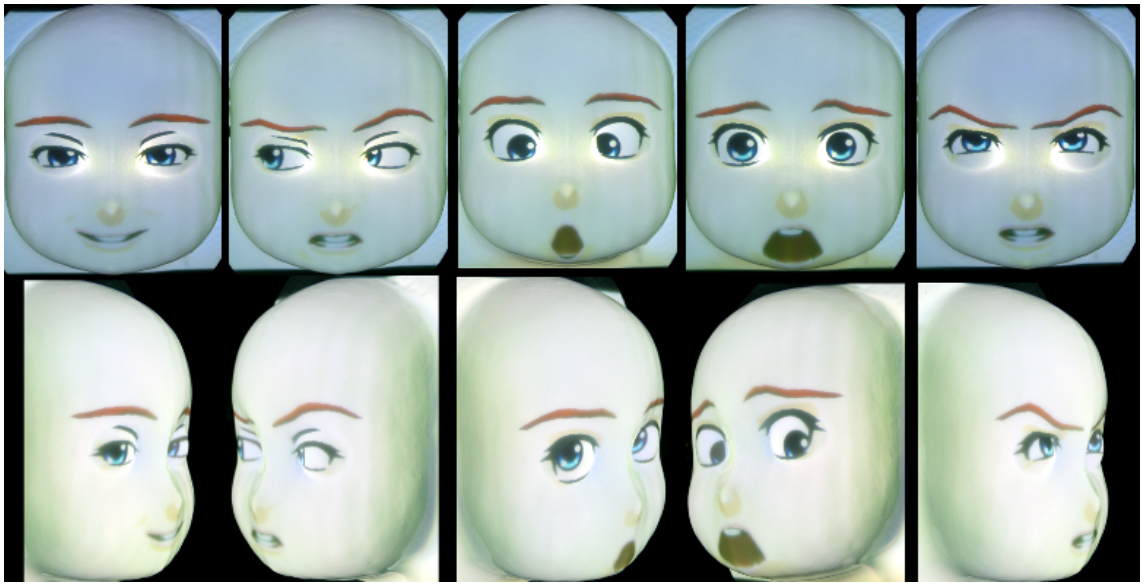


Figure 7.2: An early version of the LightHead robot face showing different facial expressions.

7.3 Gazing experiment

The LightHead robot was designed to support effective HRI. As described in section 7.1.1, the ability of interacting partners to properly interpret each others gaze is highly important. As such, for the LightHead robot to be effective, people should be able to read its gaze. To explore how the LightHead robot performs in this aspect, we conducted an experiment in which we tested how well people succeeded in reading the robot's gaze compared to other display types. This work has been published in Delaunay et al. (2010), the experiment is described in more detail below. Design of the LightHead physical face, software animations and control architecture to operate the face and robotic arm was done by Frédéric Delaunay; Joachim de Greeff and Frédéric Delaunay both contributed equally towards the experimental design, organisation, execution, analysis and dissemination of the gazing experiment.

7.3.1 Methods

To investigate how well participants can read the LightHead robot's gaze, we compare this to people's ability to read other types of faces. For this experiment four different faces were used: a real human face, a human face shown on a flatscreen monitor, a computer animated face projected in a spherical dome and a computer animated face projected in a facial mask (the LightHead robot[4]). Figure 7.3 displays the four face types, which will be referred to as human, flat, dome and mask respectively. Participants had to examine the faces from two different viewing angles, one position was frontal and the other was at an angle of 45° from the right. The resulting eight different face conditions are summarised in table 7.1.

Between the participants and the face there was a transparent grid of 50 by 50cm which was divided into 100 squares. The squares displayed the numbers 0 to 99 from top left to bottom right, so that the numbers 44, 45, 54 and 55 were in the middle (see figure 7.4 and figure 7.5 for a schematic overview of the setup). During an

[4]This last face was in fact an early prototype of the LightHead robot. When the experiment took place, the full robotic head incorporating a pico projector (figure 7.1) was not yet build. Instead, a normal projector was used to project the animated face into the 3D mask. For the objectives of the experiment this made no difference.

Description	viewing angle °	identifier
Human face	0°	human-0
Human face	45°	human-45
Human face displayed on a flatscreen monitor	0°	flat-0
Human face displayed on a flatscreen monitor	45°	flat-45
Animated face projected in a semi-sphere	0°	dome-0
Animated face projected in a semi-sphere	45°	dome-45
Animated face projected in a 3D mask	0°	mask-0
Animated face projected in a 3D mask	45°	mask-45

Table 7.1: The eight face conditions that were tested in the gazing experiment.

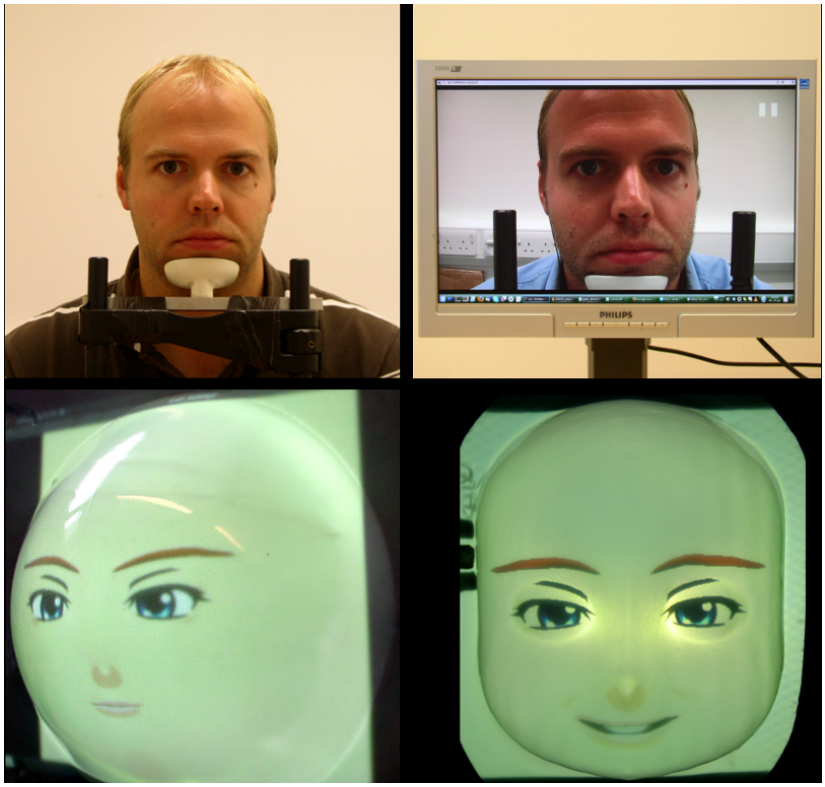


Figure 7.3: The four face types used in the experiment. From left to right and top to bottom: human, flat, dome and mask.

experimental session participants had to interpret the gaze direction of the face which they perceived through the transparent grid, and write down the number of the square they thought the gaze was directed at. A session consisted of the face looking at 50 randomly generated numbers, with 5 seconds delay between each number. In the human face condition (which was the face of one of the experimenters) the generated numbers were played in a headphone worn by the experimenter so that it could not be picked up by the participants. In the flat condition the participants were shown a pre-recorded sequence. In the case of the dome and mask conditions the sequence was generated in real-time and fed into the LightHead control system.

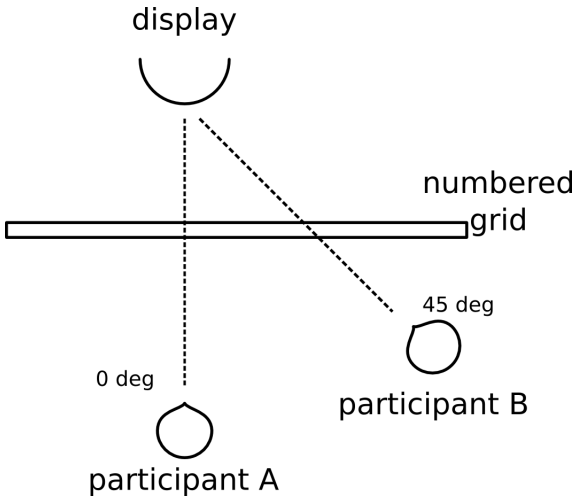


Figure 7.4: Schematic top-down view of the experimental setup illustrating the positioning of the participants, the number grid and the display showing a face.

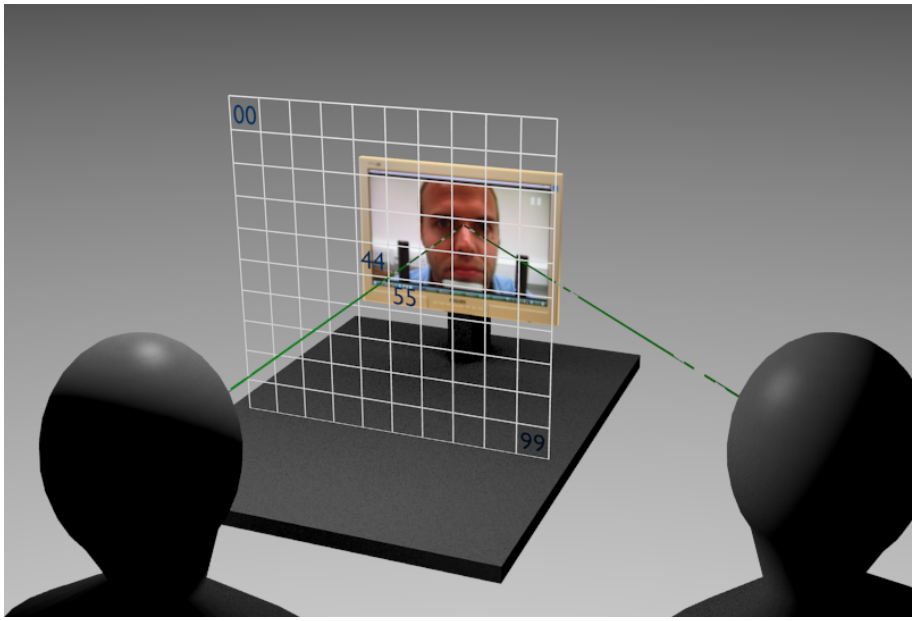


Figure 7.5: Schematic side view of the experimental setup. Subjects are facing a transparent grid with numbers, either from a 0° or for a 45° angle. Through the grid different face types can be perceived, of which the gaze needs to be interpreted.

7.3.2 Results

Each participant recorded the sequence of numbers that they thought the face was looking at. This sequence was compared to the actual sequence and the difference was calculated using the Euclidean distance, resulting in a mean gaze interpretation error for each participant. Regarding the different face types, as expected, the human face was easiest to read, followed by the mask and dome, while the flatscreen proved hardest to read (see figure 7.6 and figure 7.7).

A 4 x 2 analysis of variance (ANOVA) on gaze interpretation error showed main effects of both face type, $F(3, 88) = 8.121$, $p <$. 01, and looking angle, $F(1, 88) = 14.438$, $p < .01$. However, no interaction effects were observed, $F(3, 88) = 0.419$, $p = .740$. Comparing the different face types to each other, post-hoc comparison of the ANOVA using Tukey test revealed that the participants' performance between the human face condition and all other conditions was significant, while this was not the case for any other comparison (see table 7.2).

When examining the difference in performance between the two viewing angles, it is clear that it is much easier for participants to determine the gaze direction when they are facing the face, as opposed to a side view at 45° (error on gaze interpretation is lower for 0° than for 45°; paired samples t-test with $t(23) = -3.133$, $p = 0.005$). Comparing the performance between the two angles for the different face types, the performance difference was significant for the human, dome and mask faces, but not for flat (see table 7.3)

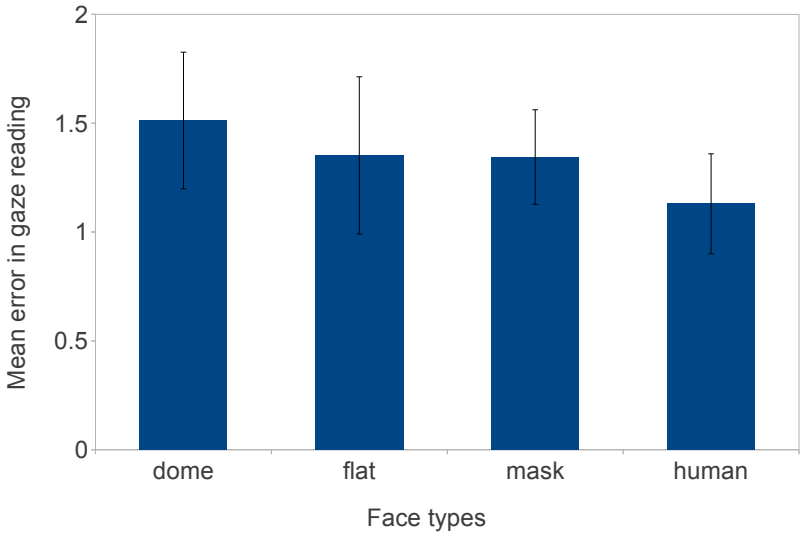


Figure 7.6: Results of the gazing experiment for the four different face types.

After the testing participants were also quizzed on their subjective experience. They were asked to indicate how effective they regarded the different face types with respect to the ability to convey gaze information, regardless of the viewing angle. This was measured using a seven-point Likert scale with the following ranges: 1-very ineffective, 2-ineffective, 3-somewhat ineffective, 4- undecided, 5-somewhat

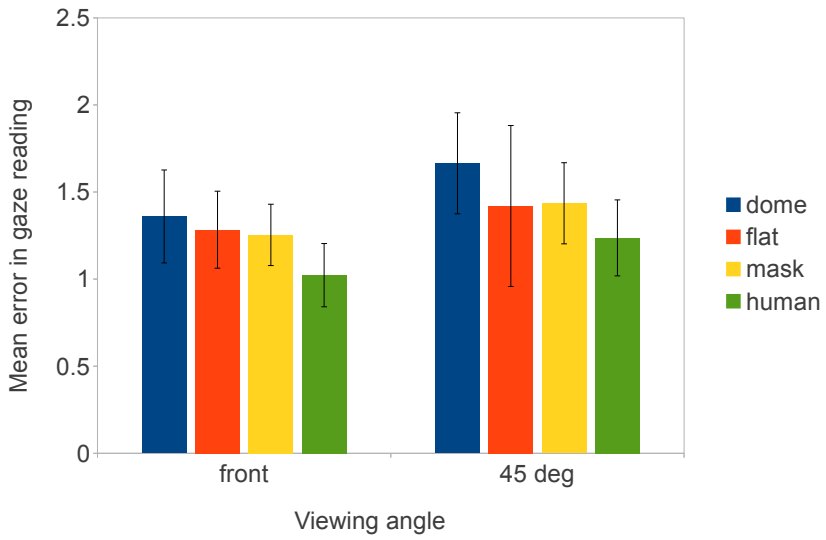


Figure 7.7: Results of the gazing experiment split into the two viewing angles and four different face types.

effective, 6-effective, 7-very effective. Unsurprisingly, the human face was rated as most effective, followed by mask, flat and dome (see figure 7.8). An ANOVA shows that the mean scores for face effectiveness were statistically significantly different ($F(2.147, 49.39) = 7.321$, $p = 0.001$). A Bonferroni post-hoc test indicated that the difference between human and dome and between human and flat was significant, while this was not the case for any of the other comparisons (see table 7.4).

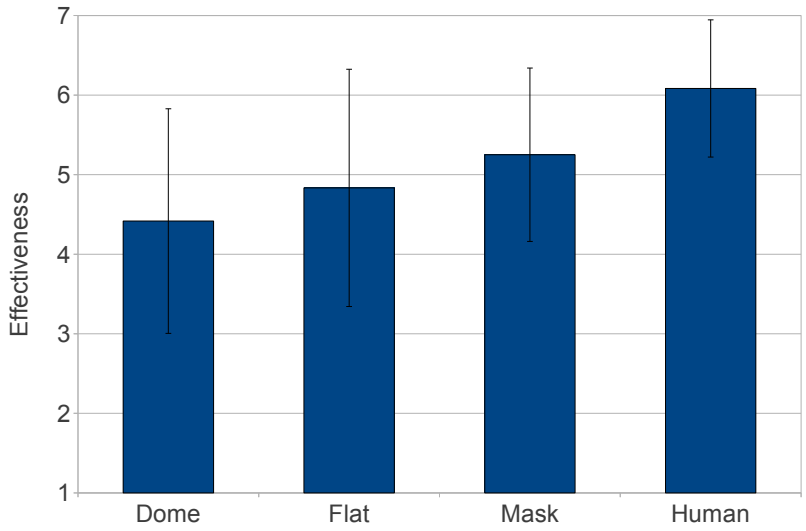


Figure 7.8: Participants' subjective rating of different face types in terms of effectiveness in conveying gaze information (seven-point Likert scale, error bars indicate standard deviation).

condition	versus	p
dome	flat	0.176
	mask	0.146
	human	0.000
flat	dome	0.176
	mask	1.000
	human	0.027
mask	dome	0.146
	flat	1.000
	human	0.035
human	dome	0.000
	flat	0.027
	mask	0.035

Table 7.2: Performance comparison of different face types. Difference in performance between human and all other conditions is significant, while this is not the case for any of the other comparisons.

	dome	flat	mask	human
t-test	t = -2.6806	t = -0.9206	t = -2.2175	t = -2.5418
	df = 22	df = 22	df = 22	df = 22
	p = 0.0137	p = 0.3673	p = 0.0372	p = 0.0186

Table 7.3: Significance tests (two-sample t-test) between the two viewing angles (0° and 45°) for the four face types. Difference in performance is significant for dome, mask and human, but not for flat.

7.3.3 Discussion

Not surprisingly, people find it easiest to read the gaze direction of another humans face. Clearly, they are most accustomed for doing this, as it is part of the daily routine while interacting with other people. However, the LightHead robot does not a bad job in terms of its ability to convey gaze direction. While it is marginally better from 0° compared to a flatscreen face, from 45° it is clearly better, as the 3D surface of the eyeballs provide more precise clues with respect to gaze direction. This is supported by the fact that compared to the dome (in which there is no 3D eyeball curvature) interpretation of the mask's gaze is significantly better interpreted (two-sample t-test with $t(22) = 2.1892$, $p = 0.0395$) from a 45° angle.

Based on verbal reports from participants it became apparent that participants not only tried to infer the gazing direction by observing the angle of the eyes, but also tried to reason about where they thought the eyes are looking. This was the

condition	versus	p
dome	flat	1.000
	mask	0.080
	human	0.001
flat	dome	1.000
	mask	1.000
	human	0.001
mask	dome	0.080
	flat	1.000
	human	0.109
human	dome	0.001
	flat	0.001
	mask	0.109

Table 7.4: Pairwise comparison of participants' preference for the different face types. Difference in preference is significant between human and dome and between human and flat, while this is not the case for any of the other comparisons.

case particularly in the flat condition at 45°. From this angle a face on the flatscreen was clearly not looking at the grid when the target was in any of the corners. Rather, the eyes appeared to be looking at some point in space that was much more on the side. However, because the participants knew that the gaze they have to interpret was supposed to be looking at the grid, they could infer that the numerical target was on the far corners of the grid when the face on the flatscreen was looking at an extreme corner. So, rather than judging the angle of the gaze based on observations alone, participants reasoned about where the gaze should be looking at, given the circumstances.

Participants also reported to find it helpful that the human face (human and flat conditions) displayed a searching strategy that was recognisable. For instance, with numerical target '66', the gaze would first vertically drop to the 6th line of the grid and then follow the numbers horizontally until 6 was found, which is a strategy that (subconsciously) might have been recognised by participants. In contrast, the animated face (dome and mask conditions) would immediately direct its gaze in the correct direction without displaying any searching behaviour.

These two aspects most likely influenced the results as they illustrate that people employed a range of strategies when interpreting gaze direction (Langton et al., 2000); the actual judgement of gaze angle is only part of this. Nevertheless, the experiment described here established that the LightHead face provides a reasonable

effective platform for HRI, as people are able to read its gaze quite well. While not on a level of human gaze interpretability yet, the 3D curvature of the eyes does improve its readability. Other cues, like searching behaviour, can be implemented quite easily due the flexibility of software animation, allowing for even more realistic gazing behaviour.

7.4 Social learning experiment

7.4.1 Background

Within developmental robotics the aspect of learning is crucial. Through effective learning mechanisms a robotic system may gain those skills that are relevant for its task. As such robots are envisioned to work in the same environment alongside humans, it would be most natural if people could teach the robot what to do. And rather than having humans, who may not be familiar and/or trained to instruct robots, adapt to the robot, it would be better if the robot could adapt to its human teacher. As such, robots might be instructed in a manner similar to how adults teach young children. To allow for this kind of teaching, a robot should be able to tap into the communication channels that come natural to people, such as speech and non-verbal behaviours like facial expressions, gestures and gaze (section 1.5).

Language and conceptual knowledge lie at the root of human intelligence, and the acquisition of both relies heavily on social interaction and tutelage (as described in section 1.5.1). Many social interactions between carers and infants are actively aimed at providing opportunities for acquiring words and their meanings, with carers overtly describing objects, actions, sensations and agents and young learners steering linguistic interactions, for example through deictic points and naming salient features in the environment. The experiment described here aims to reproduce some aspects of word and meaning acquisition in young learners, and study whether a similar mode of interacting and learning can be reproduced in human-robot interaction.

Next to this, the most natural manner of teaching is a type of interaction in

which the learner is not passively absorbing new knowledge, but rather actively engages in the learning experience. This notion of active learning was explored in chapter 5 in simulations, here the active strategies are applied to an HRI setting using the LightHead robot and human teachers. The work described here has been published in De Greeff et al. (2012b).

7.4.2 Experimental overview

In order to test the effect of an active learner, we set up a series of experiments in which the learner tries to influence the interaction with the teacher as to achieve the most optimal learning experience. This is was done in simulation first and later tested in a setup in which the LightHead robot embodied a learning agent and human subjects acted as teachers. As a learning mechanism the language game framework is used (section 3.2), and the categories of the learning agent are represented through a conceptual space (section 3.1).

To teach the robot categories, the Zoo dataset from the UCI Machine Learning Repository (Frank and Asuncion, 2010) was used, which is the same set that was used to test how the model can represent prototypes (section 4.4). The 'girl' exemplar was removed from the MAMMAL category to avoid confusion. We compared the performance of learning agents that utilises active learning (AL) to learning agents that do not (non-AL). With respect to the human teachers, we are also interested in the strategy they employ when it comes to the choice of learning examples.

7.4.3 Simulated experiment

7.4.3.1 Experimental setup

Active learning in simulation very much followed the setup that is described in chapter 5. In essence it happens as follows. During a guessing game interaction it is not the teacher but the learner that decides on the topic of the guessing game. The learner does this through examination of the context and choosing that item that is least familiar as the topic of the guessing game. As such, it allows for a quicker exploration of the conceptual space and thus yields better learning results. Note

that in the experiments reported in this chapter version AL2 is used (section 5.3). As such, the agent implements a preference to learn those items that are less familiar with respect to what it already knows. In the simulation, the teaching agent will always follow the topic chosen by the learning agent.

The agents played 50 guessing game interactions ($N_I = 50$)[5]. The context consisted of 3 animal exemplars ($C = 3$), randomly drawn from the Zoo dataset. The experiment was replicated 50 times to obtain an average measure ($R = 50$). These parameters were chosen based the aim to port this experiment to an HRI setting (section 7.4.4).

7.4.3.2 Results

We measure the ability of agents to successfully play guessing games over the course of development. The guessing success is the percentage of language game interactions in which the learner correctly identified the topic from the context based on the teacher's word. As can be seen in figure 7.9, on average the AL condition performs better than the non-AL condition, both in terms of speed (AL reaches higher guessing game success quicker), and on the long run (difference between final guessing game success). The difference in performance between the two conditions is significant (two-sample t-test with $t(98) = 8.9559$, $p < 0.001$).

7.4.4 Experimental setup of the robotic experiment

7.4.4.1 Materials

Participants were recruited around the Plymouth University campus. This resulted in a pool of 41 participants who were randomly assigned to one of the two conditions. Due to technical reasons (the robot facial projection stopped working due to overheating) two participants were removed from the pool, thus bringing the total to

[5]In a typical language game experiment agents tend to play much more guessing games (i.e. in the range of 1000 to 10,000) to more thoroughly consolidate the shared lexicon and reach communication success of about 90%. However, these numbers are infeasible to run with human participants. To be able to make a comparison between simulation and real experiment we opted for running less guessing game interactions in simulation as well. This still provides us with insights of how language games that employ AL perform in simulation.

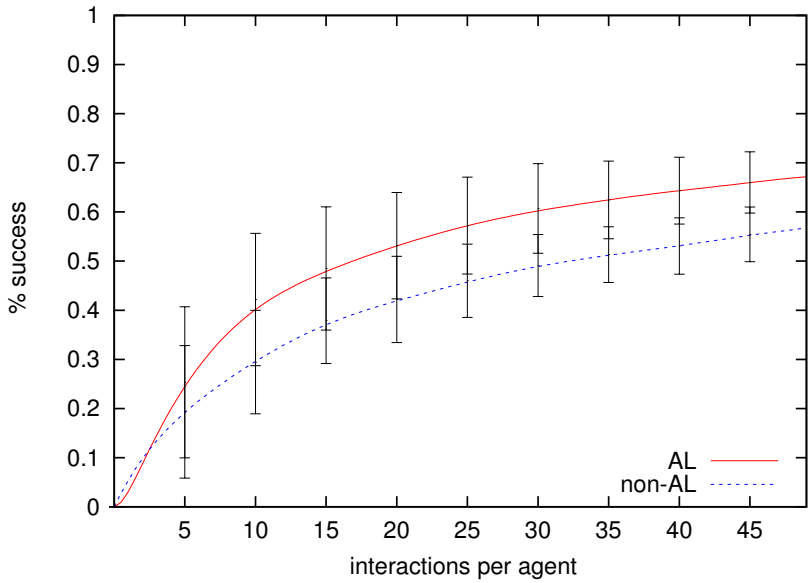


Figure 7.9: Display of guessing success in simulation for the AL and non-AL conditions.

39. The statistical breakdown in terms of native speakers, age and gender is shown in table 7.5. Participants were paid £7.50 for their participation.

	AL	non-AL	Total
number	19	20	39
native speaker	13	16	29
non-native speaker	6	4	10
female	9	11	20
male	10	9	19
age (average)	24.26	25.35	24.82

Table 7.5: Statistical breakdown of participants.

Participants interacted with the robot by means of a touchscreen. During every round of the guessing game the touchscreen displayed 3 pictures of animals along with 7 buttons to indicate animal categories. See figure 7.10 for an example. The LightHead robot was placed behind the touchscreen facing the participant and equipped with speakers to allow for speech. Different modules of the LightHead robot system that controlled facial animations, the robot speech, the robot arm and vision were run in a distributed fashion. The robot's actions were cued by interaction events picked up by the touchscreen; the camera mounted on the robot head only served to run face tracking.

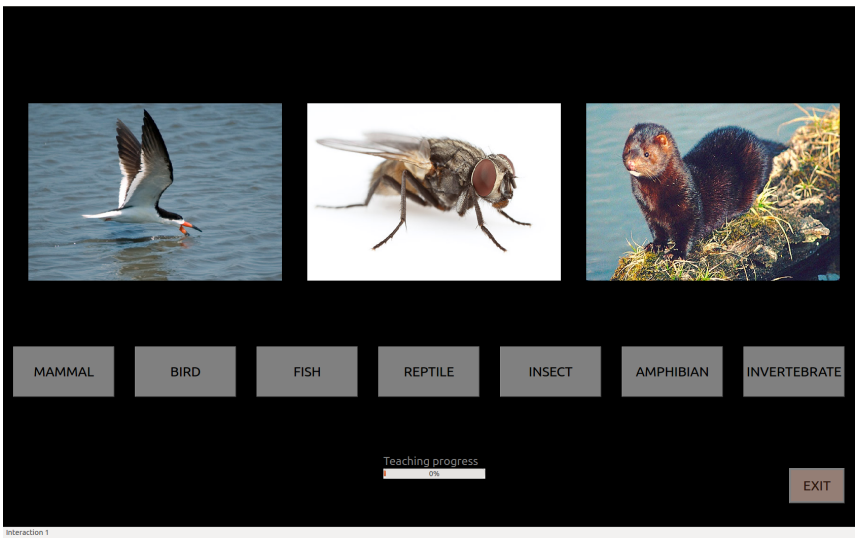


Figure 7.10: Display of the GUI which participants used to play guessing games with the robot.

7.4.4.2 Procedure

Participants were asked to sit in front of the touchscreen facing the robot. Figure 7.11 illustrates this setup. After a brief explanation by the experimenter they were invited to sit through a tutorial in which the robot explained how the guessing game was to be played. After this the robot invited the participant to play some practice rounds which involved teaching the robot colour categories. When participants were confident they knew how to play the guessing game, they could end the practice at their convenience. Occasionally the experimenter reminded participants of this option. When participants had practised sufficiently they started the teaching of animal categories by pressing a button.

The experimenter was present during the experiment, but turned sideways and occupied with other work. Occasionally participants tended to ask a questions, mostly about not knowing certain animal categories, after which the experimenter answered evasive along the lines of "just try to teach as best as you can", as not to give any clues.

The guessing game was played in a fashion similar to the one in simulation, with a human participant acting as teacher and the LightHead robot embodying the learner. During each round both the teacher and learner examined the context (3 random animal pictures displayed on the touchscreen), and depending on the

Figure 7.11: Experimental setup showing the participant, the touchscreen and the robot.

condition (AL or non-AL) the learner expressed a learning preference. The teacher mentally decided on a topic and then provided the corresponding category label by pressing the relevant button. Upon perception of the category the learner tried to guess which animal exemplar the teacher had in mind. The teacher then indicated which exemplar was the topic of the guessing game by pressing the corresponding animal picture, thus providing feedback to the learner. A more detailed description of how the robot behaved as a learner is provided in section 7.4.4.3 below. Teacher and learner played 50 guessing games.

After the experiment, participants filled in a questionnaire (appendix C) enquiring about their experience with the robot and their strategy of choosing a topic, along with a personality test. Then they were given a short debrief and given an opportunity to ask questions.

7.4.4.3 HRI

The LightHead robot acted as a physically embodied learner, which introduced some social HRI aspects. A brief description of this is provided here.

During each guessing game the robot appeared to examine the 3 animal exemplars through craning over the screen and examining each exemplar in turn. To the

participants it seemed that the robot was actually looking at the pictures[6], but this was not the case. Rather than relying on computer vision, which, given the state of the art, would not have allowed us to extract the 16 properties from the animal pictures alone, we opted for passing the properties directly to the learning agent by encoding it as a 16 dimensional vector. Instead the camera was used to run face detection, allowing the robot to follow the participant with its eyes whenever the robot was facing the participant, thus giving an impression of eye contact.

After 'examination' of the context the robot verbally invited the participant to pick a topic, without revealing the topic to the robot. Depending on the condition, the robot either moved back a bit and look at the participant (non-AL condition) or expressed its learning preference through looking back and forth from a particular exemplar to the participant while making a verbal statement such as "what about this one?" or "I would like to learn this" (AL condition). For a full list of the active learning statements the robot uttered, see table C.1 in the appendix.

When the participant had decided on the topic, he or she then pressed the corresponding category, for example MAMMAL. Upon 'hearing' this category, the robot examined the animals again and guessed which animal the participant had in mind. To express its guess, the robot fixed its gaze at an animal, and uttered an appropriate sentence, such as "is this the topic?" or "is this the animal you were thinking of?". The participants then had to click on the animal picture they had in mind, thus confirming or correcting the robot's guess. After receiving this feedback to robot expressed either joy or sadness through a facial expression and a verbal statement, depending on the outcome of the guessing game.

7.4.5 Results of the robotic experiment

7.4.5.1 Guessing game success

All participants succeeded in teaching the robot animal categories. For the final guessing success, on average the AL group was slightly more successful than the non-AL group. Final average success for AL was 0.626 ($SD = 0.077$) and for non-

[6]Which was plausible enough given the fact that the robot has a camera visibly mounted on the top of its head (see figure 7.1).

AL 0.566 ($SD = 0.087$). This difference was significant (two-sample t-test with $t(37) = 2.2917$, $p = 0.0277$). As can be observed in figure 7.12, the learning trend of both conditions is very similar to the one obtained in simulation (figure 7.9). What can also be observed from the figure is that AL speeds up learning: the slope of the AL curve is steeper and at 10 interactions the difference between AL and non-AL is significant (two-sample t-test with $t(37) = 3.6143$, $p = 0.0009$).

One aspect that may have influenced the guessing game scores is the interpretation that participants gave to the guessing game rules. As explained above, during a guessing game the learner tries to guess which item from the topic the teacher has in mind. When the learner guesses wrong, the teacher provides feedback by making it clear what was the topic. However, in the HRI setting and with the particular dataset that was used, an ambiguity can arise in those cases when two or three animal exemplars are of the same category. For instance, a context might consist of the following items: ['bear', 'lion', 'duck']. The teacher might choose 'lion' as the topic and thus would indicate MAMMAL to the robot. Upon perception of the category label and examination of the context the robot might guess the topic is 'bear' and indicate this accordingly. At this point the teacher is faced with a choice: either indicate that the topic was in fact 'lion' and thus the guessing game fails, or click on the 'bear' picture as a confirmation that, although not the topic that was originally intended, 'bear' is indeed a MAMMAL and thus the guessing game succeeds. Through informal interviews we found out that at least some of the participants gave the latter interpretation to how the guessing game should be played. Indeed, upon closer analysis of the results, it was found that some participants, in those cases in which the context constituted the ambiguity as described above, did concur with the robot's choice in a very high number of cases. That is, given a situation in which there are more than one animal exemplars of the same category and the robot guessing that one of these exemplars is the topic, the participant confirmed this by clicking on the respective animal, thus causing the guessing game to succeed.

We have no measure to determine to what extend participants might have chosen a certain animal as the topic and then acknowledging the robots guess when it picked

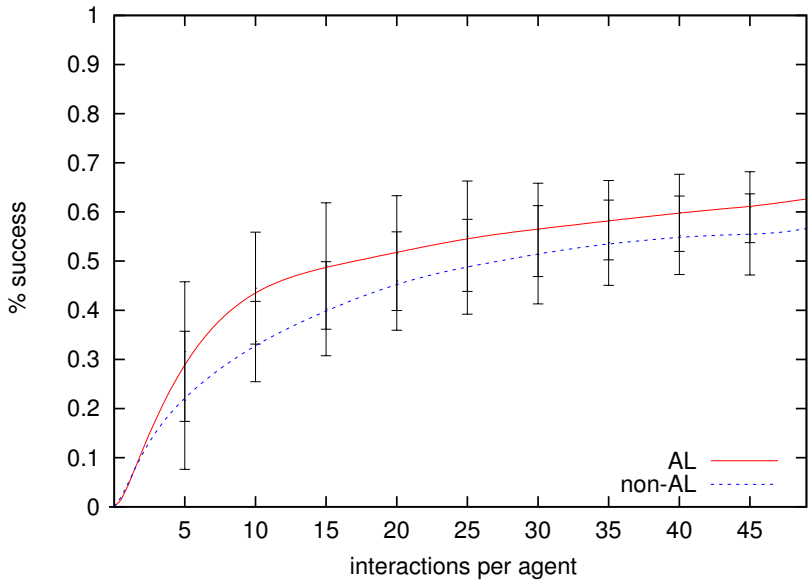


Figure 7.12: Display of guessing success from the AL and non-AL groups.

a different animal but of the correct category. To gain some insights in how often this might have happened, we measured for each participant the number instances during which two animal exemplars were from the same category. We then counted how often participants confirmed the guess the robot made regarding the topic of the guessing game for these instances. Dividing this number by the total number of instances during which the potential ambiguity arose gives a percentage for each participant. As can be seen in table C.2, for some participants this percentage is quite high. Thus, it seems justified to assume that indeed, some participants interpreted the guessing game as described above.

7.4.5.2 Response to active learning

To get an idea of how much participants responded to the active learning behaviour of the robot, we measured the proportion in which the robot's preferred topic was similar to the one the participant indicated they had chosen as the topic. For non-AL this is 0.32 (SD = 0.08), as the robot does not provide social cues and hence the topic choice is random. For AL however, this proportion turned out to be 0.56 (SD = 0.18), indicating that the level at which participants followed the robot's choice was more than chance ($p < 0.001$). Thus, on average, participants did respond to the robot's social cues. The individual scores for all participants is shown in

table C.3. What can be seen clearly is that there are quite some individual differences amongst the AL group; some participants completely ignored the robot's social cues, while others strongly responded to this (most notably participant #8 with 94% and participant #37 with 86%, table C.3). Figure 7.13 plots the responsiveness to AL against the final success rate of the teaching. Because of this high variance only a weak correlation between the participants' responsiveness to AL and the guessing game success rate was found (AL condition, Pearson's $r = 0.09$).

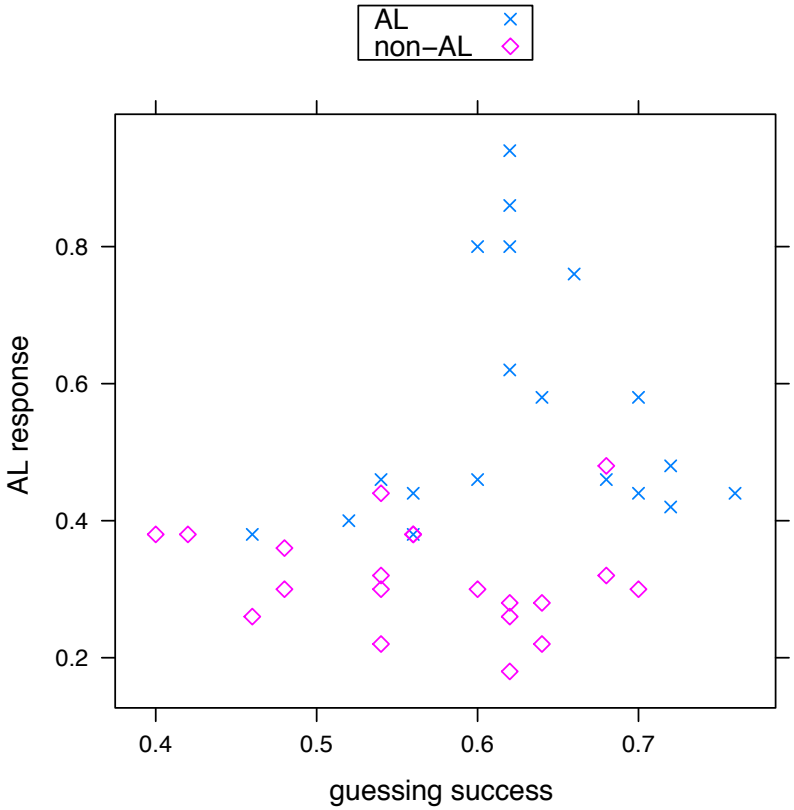


Figure 7.13: Distribution of responsiveness to social cues against learning performance for AL and non-AL groups. In the AL condition, the robot provided social cues, which were picked up by some participants and had a positive effect on the robot's performance.

7.4.5.3 Category use

Because teachers are free to choose the topic for each guessing game as they see fit[7], the distribution of category use, i.e. the frequency with which the different categories are chosen as topic for the guessing game, is of interest; this may reveal some

[7]In simulations this is governed by either AL mechanisms or through random choice.

aspects of the teaching strategy (if any) that is used by the teacher. A normalised distribution of category use is derived by dividing the total uses of each category for all participants from the same condition by the total number of games played. For instance, in the AL condition, the category MAMMAL was used 261 times in total by all participants from this condition. This number is divided by 950, the total number of guessing games that was played in the AL condition (19 participants, 50 games each), resulting in a normalised use of 27.5%. Figure 7.14 displays these normalised distributions for the non-AL and AL conditions and for the actual distribution of the dataset.

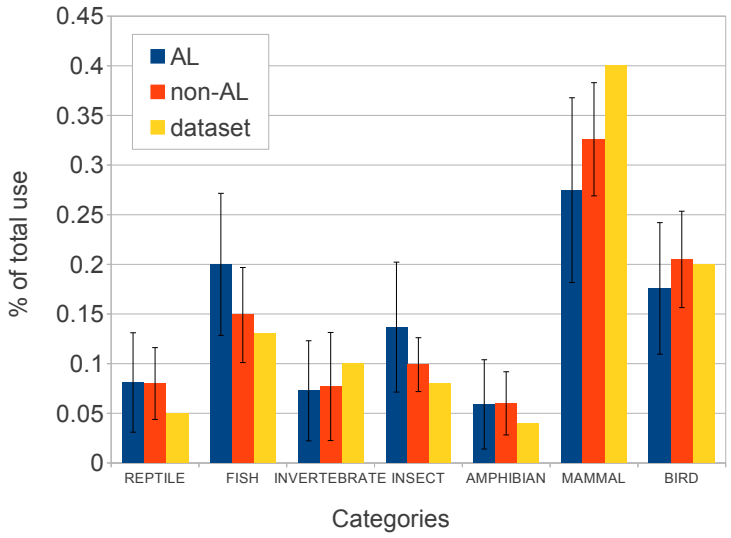


Figure 7.14: Normalised distribution of category use in AL and non-AL condition compared to the distribution of the dataset.

Because the AL condition constitutes a more optimal use of categories with respect to learning the dataset, that is, AL results in higher communicative success, it is interesting to see that participants in the AL condition diverge from the distribution of the dataset. The non-AL group is somewhat halfway, indicating that participants in this condition still follow a strategy rather than picking topics at random. The tendency to diverge from the dataset distribution can be seen as an indication that participants actively try to provide specific learning examples, i.e. their choice is not random (indeed, when asked for this, participants reported various reasons, see section 7.4.5.4). If all categories were used equally, each category

would be used for 14.3% of the cases. If we calculate the difference between category use and this uniform use for each group, we see that the sum of differences is 0.44 for AL, 0.50 for non-AL and 0.63 for the dataset distribution, indicating that AL is closest to equal category use, followed by non-AL and the dataset distribution is the least equal.

7.4.5.4 Questionnaire

As described in section 7.1.2, a questionnaire is a common method to gain insights in participants' subjective experiences of an experiment. As such, after the participants were done with the teaching of the robot, they filled in a questionnaire (appendix C) asking them to rate the following questions on a seven-point Likert scale:

Q1: How do you rate your interaction with the robot?
Range: [not satisfactory at all - very satisfactory]

Q2: How do you rate the robot's behaviour?
Range: [not natural at all - very natural]

Q3: Do you have any experience with robots?
Range: [I have no experience with robots - I have a lot of experience with robots]

Q4: Who was in control of the teaching sessions?
Range: [I was in control - the robot was in control]

Q5: On what basis did you choose the animal examples as topic? Please explain.

Q6: Do you like science fiction (books, film, etc)?
Range: [I don't like science fiction at all - I very much like science fiction]

Q7: How many emotions do you think the robot has?
Range: [the robot has no emotions - the robot has a lot of emotions]

Q8: How smart do you think the robot is?
Range: [the robot is not smart at all - the robot is very smart]

Q9: How many hours per week do you spend using a computer?

Q10: General comments

The mean responses to these questions (except questions 5 as this is a qualitative question) split into AL and non-AL groups do not show any significant differences. Most results are what was expected, e.g. for Q1 participants rate their interaction a bit higher in the non-AL condition, presumably because they are more in control. For Q2 they rate the robot's behaviour a bit higher in the AL condition, possibly because the robot is more vocal here. The robot was also perceived to be somewhat smarter (Q8). Contrary to what was expected, no difference in the perception of control was found (Q4). Also, participants attribute more emotions to the robot in the non-AL condition (question 7; however, the difference is not significant, two-sample t-test with $t(37) = -1.7943$, $p = 0.081$). When gender is also taken into account, some interesting patterns emerge. These are discussed in more detail below (see figures C.1 and C.2 for all figures). The answers to question 5, which asked participants on what basis they chose an animal exemplar as topic for the guessing game varied widely. The following are some examples of the answers participants gave: "What I thought would be easy for the robot to learn", "Often where the robot looked to, what seemed to be the easiest for me", "Which animal I prefered, how cute & fluffy it looked, or how interesting I thought it was". For the full list of answers, see appendix C.1.1.

A correlation test between the guessing success and all questions is displayed in table 7.6. Weak negative correlations exist between guessing game success and Q6 and Q7; a weak positive correlation exist between guessing success and Q8. Between the responsiveness to AL and the answer to Q4 there exist a weak positive correlation (AL group, Pearson's $r = 0.24$).

7.4.5.5 Gender differences

When examining the guessing game results for gender differences, an interaction can be found between active learning condition and gender. It appears that in the

	Pearson's r
Q 1	0.041
Q 2	0.119
Q 3	-0.157
Q 4	-0.059
Q 6	-0.400
Q 7	-0.380
Q 8	0.280
Q 9	-0.154

Table 7.6: Correlation test between guessing success and participants' questionnaire answers.

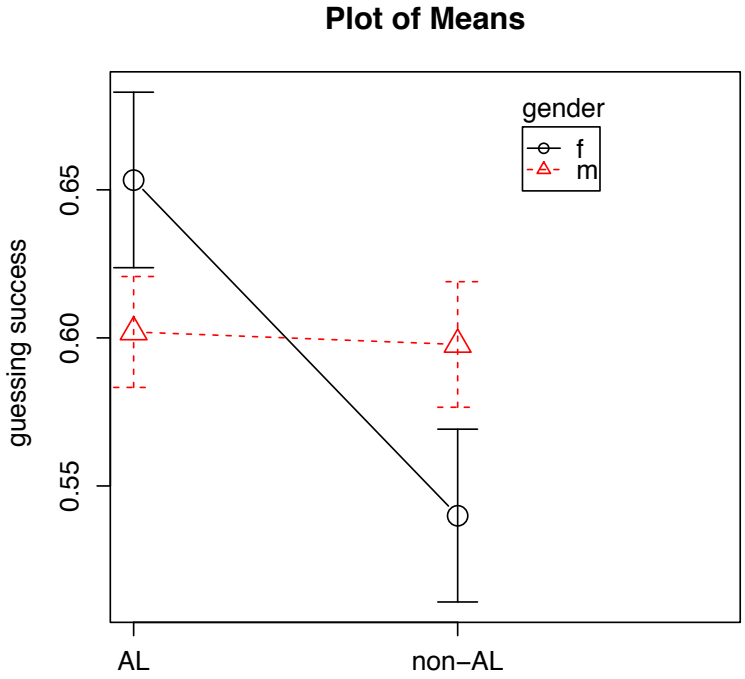


Figure 7.15: Guessing success split into AL/non-AL and gender.

case of female teachers the robot is more successful in guessing games in the active learning condition than with male teachers (see figure 7.15). Running an ANOVA indicates that the interaction is significant with $F(1, 35) = 4.5697$, $p = 0.0396$.

Another aspect that indicates gender differences was found in the analysis of the questionnaire. Particularly Q2, Q4 and Q8 are interesting in this regard, as they express the participants' attitude towards the robot. The corresponding figures split in gender and gender & AL/non-AL are displayed in figures 7.16 to 7.18, an analysis is provided below.

- Question 2:"How do you rate the robot's behaviour?" As can be observed in the figure, there is a clear difference in rating between genders, as females rate the robot's behaviour significantly higher in the AL condition (figure 7.16, right; an ANOVA shows interaction with $F(1, 35) = 8.517$, $p = 0.006$).

- Question 4: "Who was in control of the teaching sessions?" The working assumption has been that in the case of AL the robot is more explicit in its learning preference, and therefore participants would perceive the robot to be more in control. This is not what was observed. However, if the response to question 4 is split by gender, we can observe that on average female participants judge the robot to be much more in control than male participants (figure 7.17, left; two-sample t-test with $t(37) = 3.2805$, $p = 0.0023$). When the AL/non-AL factor is included as well, opposing patterns can be observed (figure 7.17, right). No significant interaction was found (ANOVA with $F(1, 35) = 1.814$, $p = 0.187$), but a trend can be observed.

- Question 8:"How smart do you think the robot is?" The difference between AL and non-AL is not significant (two-sample t-test with $t(37) = 1.6366$, $p = 0.1102$), but it does indicate a trend. Looking at the relation between AL/non-AL and gender, significant interaction is found (ANOVA with $F(1, 35) = 6.229$, $p = 0.017$), clearly indicating that female participants consider the robot to be smarter when it displays active learning behaviour (figure 7.18, right).

7.4.5.6 Personality test

Along with the questionnaire participants were asked to fill in a personality test based on the Big Five Inventory (John et al., 2008). This test is very common in psychology; rather then classifying people into predefined personality types (as happens in the Myers-Briggs Type Indicator) people are scored on 5 dimensions (openness, conscientiousness, extraversion, agreeableness, neuroticism) that together provide an indication of their personality. The main purpose of this was to be able to check whether or not any outcome from the teaching experiment (e.g. responsiveness to active learning) could be explained through certain personality traits.

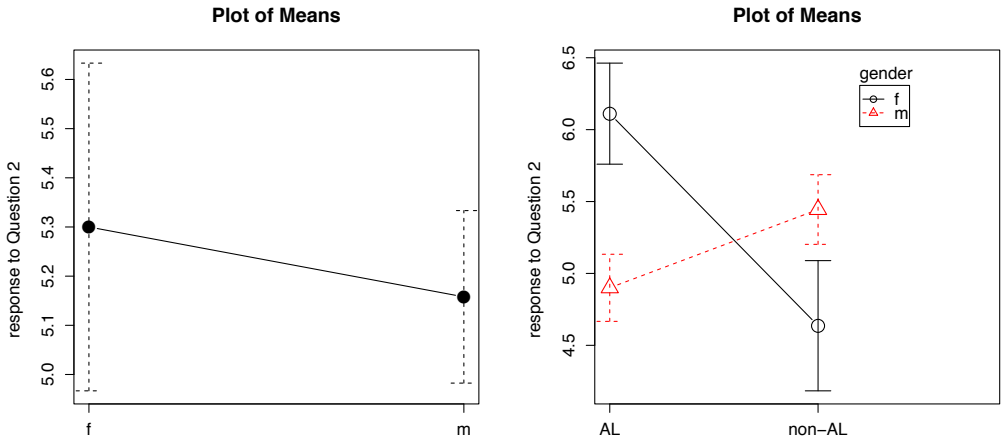


Figure 7.16: Response to question 2 "How do you rate the robot's behaviour?" split by gender (left) and by both gender and AL/non-AL (right).

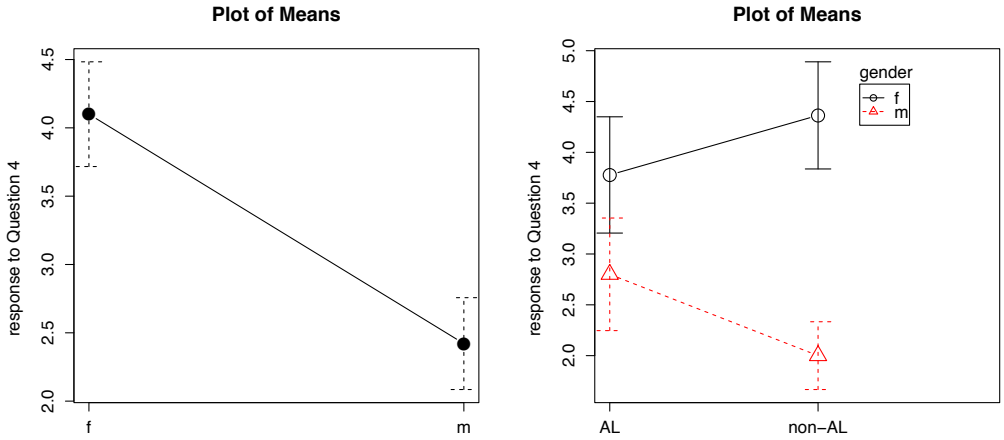


Figure 7.17: Response to question 4 "Who was in control of the teaching sessions?" split by gender (left) and by both gender and AL/non-AL (right).

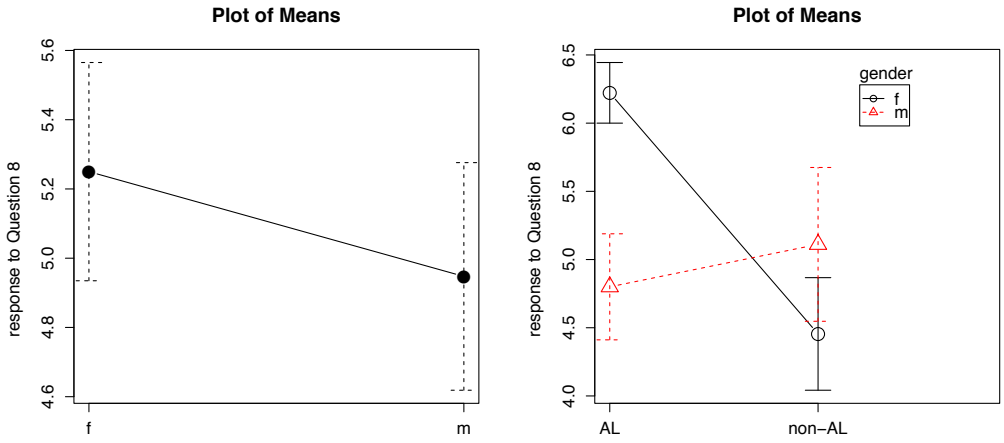


Figure 7.18: Response to question 8 "How smart do you think the robot is?" split by gender (left) and by both gender and AL/non-AL (right).

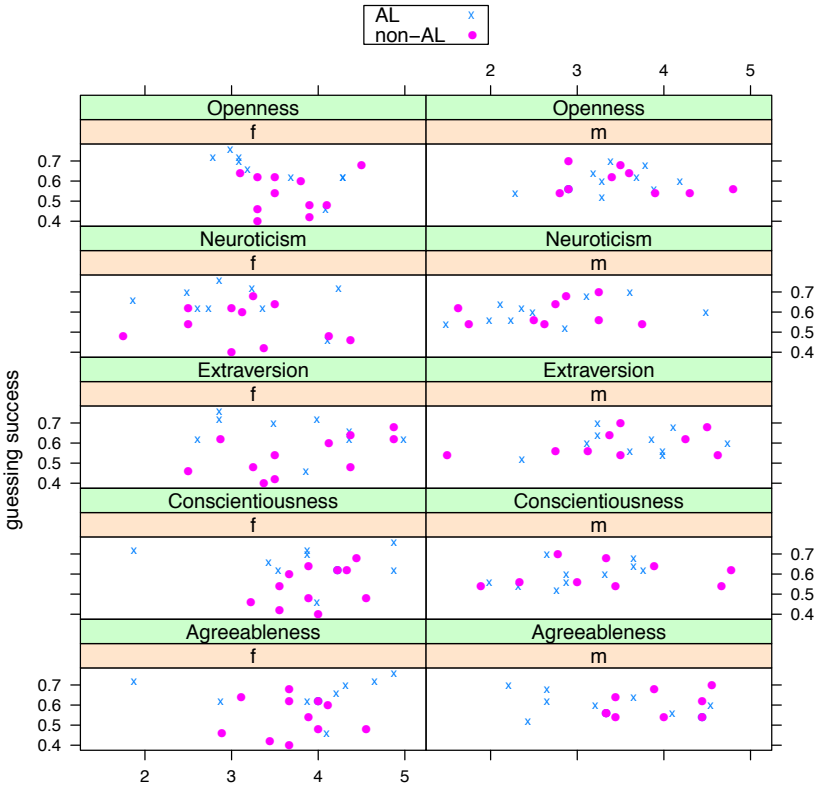


Figure 7.19: Personality trait scores split into AL/non-AL and gender, against guessing success.

A medium correlation was found between the responsiveness to AL (AL condition only) and the conscientiousness trait (Pearson's r = 0.424, not significant with $p = 0.071$). Furthermore, the difference in conscientiousness between female and male participants was significant (two-sample t-test with $t(37) = 3.1488$, $p = 0.0032$). Both these findings fit in with the gender differences described above, which indicated that females are more responsive to an active learning robot. No other influences of personality traits on any of the other outcomes were found, and as such it was concluded that different personality traits do not appear to influence the effectiveness of active learning by a large amount. Figure 7.19 displays all personality trait scores split into gender and AL/non-AL against guessing success.

7.4.6 Discussion

The ability for a robot to learn from a human teacher is deemed important for achieving effective robot systems that can co-exist with humans in an unknown

environment. Furthermore, to utilise learning experiences to their fullest potential, an active learner may be able to shape the learning interaction in such a way as to experience the most effective learning.

This section has presented experiments in which an active learner (in simulation and embodied in the LightHead robot) is able to positively influence the learning experience offered by a teacher through utilisation of social clues. This allowed the learner to learn quicker and more effectively, as illustrated by an improved learning curve depicting guessing game success. Furthermore, when the learner is embodied in a robot, it was shown that participants are responsive to the social clues displayed by the robot, albeit to varying degree. Participants interacting with an active learning robot are more likely to choose as learning examples those categories that are closer to an optimal distribution of category use, compared to the participants that interacted with a non-active learning robot. Analysis of the experiment and responses given to a questionnaire revealed gender differences, indicating that female teachers were more receptive to the social clues of an active learning robot. This suggests that there is no optimal strategy that works best for all human tutors, Rather, the most effective learning experience might be achieved by a learner that personalises its behaviour towards individual teachers, as different teachers may respond differently to social cues.

7.5 Chapter summary

In this chapter experiments were extended to a real life study through the inclusion of a robot learner and human teachers. The inclusion of robotic hardware constitutes a more thorough test environment compared to a simulation-only approach, as a real environment in which robots and people interact demands more robustness. The LightHead robot was discussed, which is a new kind of robot specifically targeted at HRI and which offers benefits over more classic approaches due to the utilisation of back-projected animations. An experiment in which participants were tested for their ability to read the robots gaze indicated that the LightHead face is quite suitable for this, strengthening the use of this robot for HRI purposes, as gaze

reading is an important social cue.

The LightHead robot was then used in experiments that extend the experiments as reported in chapter 5 in an incremental fashion, i.e. the assessment of active learning strategies into a real life setting. In this, human participants acted as tutors for the robot which utilised active strategies. It was shown that these active strategies were still effective in a real world setting, although the best results might be achieved when the robot learner's behaviour is even more tailored towards individual teachers. These experiments illustrates that people are willing to engage in a tutoring relation with a learning robot and moreover that an active robot learner is able to shape the learning experience to achieve better results through utilisation of social behaviour.

Chapter 8

Summary, discussion and future work

This chapter provides an overview of the topics covered in this thesis as well as concluding remarks. Main themes are recapitulated and reflected upon, contributions and shortcomings are discussed and some suggestions for future work are provided.

8.1 Summary

This thesis has investigated a range of aspects related to social learning of concepts by artificial agents, in which agents were both simulated and embodied in physical robotic hardware. The ability to use concepts as bearers of knowledge has been identified as a crucial aspect of human cognition; as such, artificial systems that are intended to operate on a level comparable to humans will need to be able to utilise conceptual structures in similar vein. To achieve this aim various routes can be considered; by now it has been fairly well established within cognitive systems and AI research that an intelligent system cannot be considered in isolation. Rather, a more holistic approach which takes into account the environment, the embodiment and the interactions that a system might have with its social and physical surroundings, is perceived as more viable. Particularly the learning of words and concepts has been recognised as a two-way process in which both teacher and learner mutually contribute to the learning experience.

After reviewing various psychological theories with respect to human concept use, the standpoint has been adopted that a functional model of concepts should at least be able to represent concepts as prototypes. Indeed, a model in which prototype concepts function as abstract summary representations is considered adequate, particularly in light of the ease with which features from exemplar theory might be incorporated, thus combining two of the main theories from the conceptual literature. While prototype and exemplar models are not the 'final' solution to conceptual modelling, as more advanced hybrid models do exist, the former were nevertheless considered sufficiently adequate with respect to the aims of this work, and the implementation of these theories through the conceptual spaces framework can be considered as a minimalistic model of concept representations. As this work is not necessarily aimed at providing a conceptual model that maximises explanation of empirical data from human concept use, the adoption of a more minimalistic model is not viewed as a drawback, but could be considered a virtue in light of notions like Occam's razor.

The conceptual spaces framework and the language game framework were adopted as tools to model social concept learning. A model based on conceptual spaces provides a means of conceptual representation compatible with prototype theory; adequate functioning of the CS model was established through an exploratory study involving the formation of prototypes based on a commonly used dataset of zoo animals. The application of language game dynamics emphasises the interactive aspects of learning; categories, consisting of word labels associated with points or regions in the CS, become shared within a population of agents through repeated linguistic interaction. In this approach, an agent learns the meaning of concepts through interaction with its environment which includes other agents; the model as such incorporates social and interactive aspects of concept learning.

Using the model as described above, the notion of active learning was further explored in a series of experiments based in simulation. In this, the learning strategies of an agent are augmented with some more elaborated mechanisms which enable the agent to influence its learning experience. These mechanisms were active learning,

through which the agent can influence the dynamics of the language game as to explore unknown stimuli; knowledge querying, which allows the agent to query its teacher on less well established concepts; and contrastive learning, which contrasts just learned concepts with other examples as to more firmly establish the association between words and objects. Of these three mechanisms, active learning turns out to be the most effective, resulting in a small, yet significant improvement with respect to the speed and the quality of learning.

Subsequently the role of embodiment was explored, in particular the effect of difference in embodied perception on the formation of concepts. A remarkable phenomenon is the large degree of agreement expressed by people regarding colour concepts, despite the fact that there exist relatively large physiological differences in the retinas of human individuals. This agreement amongst people with respect to colour naming has led researches to suggest that "neurological factors" play a role in the stabilisation of colour names; however, how these neurological factors function is relatively less explored. Running experiments with the model as described above allowed for the exploration of the hypothesis that this agreement can be explained through a dynamic interaction in which agents, that on a low level perceive the world differently, negotiate and align shared vocabulary to describe colours. This resulted in effective communication systems despite agents' perceptual differences in both software simulations and on robotic hardware. We suggested that the application of similar dynamics may provide an answer to the question of how people overcome the stark differences in their cone-ratios while simultaneously being able to communicate about colours without much problems.

Taking the notions of social, embodied and active learning of concepts to a real-life situation requires the use of a robot. The LightHead robot constitutes a novel robotic platform which was specifically developed with HRI in mind. The robot sports retro-projected animated face technology which allows for a wide range of facial expressions, as such providing a flexible channel for non-verbal communication. Given the importance of mutual gaze understanding in social interaction, it is considered paramount that people interacting with the LightHead robot are able

to interpret its gaze correctly. To verify this, experiments were conducted in which participants had to read the robot's gaze; this was compared to other face types and led to the conclusion that the LightHead's gaze can be adequately interpreted, establishing the robot as a functional platform for further HRI experiments. The robot was then used in an HRI experiment in which an agent, embodied in the robot, employs the active learning strategies as described above. In this experiment a human teacher taught the robot concepts based on animal classifications. Utilising both verbal and non-verbal channels, it was established that the active learning strategies can also be beneficial in an HRI setting.

8.2 Discussion

The models and experiments presented in this work are believed to provide a reasonable approach to interactive social learning of concepts by embodied agents. Nevertheless some remarks and points of discussion can be made. These issues are discussed in this section.

Advances in science The fact that conceptual structures play a crucial role in human cognition is considered a given; however, it is relatively less well established what concepts are exactly, which functional role they play in various cognitive operations and how they intertwine with other aspects of cognition such as memory, reasoning and perception. This is an ongoing debate and research agenda within psychology, neuroscience and philosophy. Recent advances particularly in neuroscience have resulted in increased understanding of the fine grained operation of the brain; however, it remains clear that this is an ongoing research endeavour as currently the field is relatively young and a lot of the inner workings of the brain are not properly understood yet. It is expected that advances in these fields will reflect on psychological, neurological and computational models of concepts, as neurological insights allow for more advanced cognitive models that might better explain empirical data. These models will subsequently find their way to the more AI and robotic oriented applications.

Choice of models Even though the choice of a conceptual model compatible with prototype/exemplar theory is based on psychological insights and as such is positioned within empirical data; other, potentially more advanced, models of concepts are also conceivable. While the models used within this work possess adequate characteristics that support the learning of concepts in interactive embodied agents, it could nevertheless be argued that these models can not necessarily account for all aspects of concept modelling. For instance, two crucial aspects of concepts, hierarchy and compositionality, are not accounted for within the frameworks that were used. This could be considered a shortcoming if the aim of the model is to be compatible with more (if not all) empirical data regarding human concept use. These aspects are discussed in more detail in section 8.3.1. Another point that could be made is the fact that the models used in this work are not necessarily biologically plausible, as the manner in which prototypes are modelled in a conceptual space is fairly abstract. This is not strictly speaking incompatible with biological plausibility; however other, more neurologically inspired accounts that more closely model the fine-grained operations of the brain might provide explanations that are closer to actual biological systems.

Related to choice of models are the algorithmic and computational choices that were made for the work described in this thesis. While conceptual spaces provide adequate means to model prototypical structures (as illustrated in section 4.3), other techniques exist that can provide similar means of segregating and structuring sensory input space. The language game model provides a reasonable mechanism of associating perceptual data with linguistic labels; it was chosen particularly because of the emphasis it places on social and interactive aspects of learning. However, in similar vein, other techniques exist that can provide comparable means of learning associations between different modalities and linguistic labels. Fairly common are Hebbian-like associative mechanisms, for which it has been argued that they resemble learning in natural systems, e.g. (O'Reilly, 1998; Munakata and Pfaffly, 2004). Indeed, such associative mechanisms are employed by the ERA (Morse et al., 2010) and the DAIM (Baxter et al., 2011, 2012) cognitive architectures. Overall the view

is endorsed that a cognitive system needs two components; the first one is a means to segregate perceptual data streams into manageable and meaningful 'chunks', i.e. to create some kind of order in the "blooming, buzzing confusion" (James, 1890, p. 488); the second component is a mechanism to form associations between different chunks of perceptual data, potentially governed through language. Whether language is determining, guiding or merely influencing the formation of associations between perceptual structures is still a topic of debate (as discussed in section 2.3.1), but it remains clear that language has a special role in the development and operation of human cognition.

Abstraction The concepts that have been modelled in this work are relatively concrete, e.g. colours, shapes and animals, and typically have observable properties[1]. In contrast, for more abstract concepts such as UNEMPLOYMENT, ETHICS or EVOLUTION it is much harder to determine the underlying properties on which these concepts are based. As concepts are to be represented in a conceptual space, it is vital that the relevant dimensions that are to form the axes on which the concept is expressed are available. Particularly for more abstract concepts this is not always that straightforward. Gärdenfors (2000b, p.21) suggests this might be done in conceptual spaces through techniques like multidimensional scaling (MDS), which requires similarity judgement data from people regarding a set of (abstract) concepts. The underlying dimensional space can then be constructed through application of the MDS algorithm. However, as Gärdenfors acknowledges, it might be hard to give a meaningful psychological interpretation for the resulting dimensions.

Various authors have suggested that abstract concepts are somehow based on lower level cognitive capacities, as such obtaining a hierarchical structure which reflects a continuum from concrete to abstract (Wiemer-Hastings et al., 2001). In this view abstract concepts are envisioned to be based on metaphors (Lakoff, 2008), simulations (Barsalou, 1999) and/or actions (Glenberg and Kaschak, 2002). Embodiment plays an important role in this, as bodily constraints influence the manner in which

[1]The properties of the animal exemplars that were used in the experiments are provided by the dataset; that is, the dataset dictates which animal has which properties, this does not need to be inferred from visual appearance by e.g. computer vision algorithms.

these metaphors, simulations and actions can give rise to abstraction (Barsalou, 2008). Related approaches have tailored their proposed solutions towards developmental robotics, e.g. Stramandinoli et al. (2012) illustrate how higher-order concepts might be indirectly grounded in action primitives directly grounded in sensorimotor experiences of the iCub robot. In similar manner, it is conceivable that concepts expressed in conceptual spaces with relative concrete dimensions could be invoked to play a role in the representation of more abstract concepts, e.g. as building blocks, metaphors or simulations.

Statistical learning While social and interactive concept learning appears to be an effective approach with respect to having a robot learn concepts (as illustrated by the research reported in this work), it is nevertheless not the only manner in which an artificial system might learn. Other approaches that do not rely on social and supervised means have been relatively successful. These approaches are mostly based on aggregation of statistical regularities of the environment in which a system operates. Examples of these approaches are latent semantic analysis (as discussed in section 2.4.3), which allows for the formation of a coherent network of semantics purely though exposing the system to large text corpora.

Another approach is cross-situational learning, which allows a learner to acquire semantic knowledge in situations where there exist perceptual ambiguity regarding the referent of a word, i.e. situations in which it is unclear what exactly is meant with a word label (Belpaeme and Morse, 2012). While compared to social learning through e.g. language games, cross-situational learning mechanisms perform less well in terms of speed and performance, they do represent additional means in which a system might acquire semantic knowledge, as such solving referential uncertainty. It has been shown that such strategies are employed by both children and adults by Smith and Yu (2008) and Smith et al. (2011) respectively.

Experiments that combine both social learning mechanisms and mechanisms based on statistical regularities are presented in (Vogt and Divina, 2007; Vogt and Haasdijk, 2010). These experiments investigate the relative influence of various learning mechanisms such as feedback, the principle of contrast, joint attention and

cross-situational learning on the ability of a population of agents to develop a shared language. It is concluded that, under varying conditions, both social learning mechanisms (such as joint attention) and mechanisms of learning based on statistical regularities (such as cross-situational learning) contribute to the development of a successful communication system in a population of agents. As such, a hybrid approach in which both social learning mechanisms and strategies based on statistical regularities (both cross-situational learning which is based on statistical occurrence of word labels in a context provided by a (simulated) environment, but also latent semantic analysis which relies on statistical occurrence of words within the context of other words) could constitute a promising approach.

Social learning The notion of social learning applied to cognitive robotics, or socially guided machine learning, is a relative new research agenda that has not fully matured yet. Recent advances, as discussed in section 1.5.1, have shown the feasibility of this approach, indicating that there is much to gain when learning artificial systems are able to utilise social channels that are, in a sense, readily available when interacting with people. The reason that this field is relatively unexplored is mainly due to the fact that only in the last decade or so robots have become sufficiently advanced in both their cognitive capabilities and appearances that they can be perceived as 'natural' social partners by people. Examples of such robots are for instance: Kismet, iCub, Asimo, Nao and LightHead. Many more exist, see Fong et al. (2003) for a survey.. This is a prerequisite for social learning, as crucial to this strategy is the fact that people interacting with a robot should be able to do so without any specialised training.

The ability of robots to behave socially appropriate is fundamental, as this can tap into the teaching, guiding and tutoring behaviours that people naturally possess (Breazeal et al., 2004; Weiss et al., 2010). Given the relative novelty of this approach in combination with the promising results so far, it is expected that social learning will have a growing impact on future applications of cognitive robotics. Indeed, it appears that most research directed at this topic so far has been of pioneering nature and as such much more is to be expected. In light of the advances in behaviour,

capacities and appearances of cognitive robotics, advances in understanding of social intelligence in humans, and suggestions that robotic learning behaviour personalised towards individual tutors appears to be even more effective, interactive learning robots are envisioned to be a very promising approach.

8.3 Future work

In this section we describe some topics that are perceived to be valuable additions to models dealing with social learning of concepts. Support for the first topic, hierarchy and compositionality, would result in richer conceptual structures; addition of associative networks and statistical forms of learning would enable the model to exhibit features such as semantic priming; and thirdly, further utilisation of active learning through HRI could greatly enhance a robot's learning effectiveness. These topics constitute future work and are discussed in more detail below.

8.3.1 Hierarchy and compositionality

While studying the literature and previous work on the topic of concepts, it has become apparent that two specific properties are important for concepts. These properties are *hierarchy* and *compositionality*. Both are considered as fundamental of concepts, and any theory or framework that is to capture the manner in which concepts are used by people will have to address these.

Hierarchy This refers to the fact that concepts are typically part of a hierarchical network consisting of different levels of description (Murphy and Lassaline, 1997). For example, are 3-level hierarchical network may consists of higher level concepts (superordinates) that form a more general level of description, medium level (basic level) concepts that form a 'regular' version of a concept and a lower level (subordinate) concepts that are the most specific version; a typical example is: ANIMAL → DOG → LABRADOR. Basic level concepts (DOG) have special status; they strike a balance between generality and specificity, they are learned earlier and are generally fastest in retrieval tasks (Rosch et al., 1976; Mervis and Crisafi, 1982; Murphy

and Brownell, 1985); although this basic-level advantage can be affected by domain specific knowledge of individuals (Tanaka and Taylor, 1991). Hierarchical levels of representation have been implemented within language games by Vogt (2004). In this work it was illustrated how the use of different levels of generality regarding concept representation (and an incentive for an agent to generalise categorisation as much as possible) can lead to the emergence of the Zipf-Mandelbrot law.

Compositionality Compositionality refers to the fact that simple concepts can be concatenated to form more complex ones. For example, the simple concepts FIRE and TRUCK can form the more complex concept FIRETRUCK. Because FIRETRUCK is evidently something else than simply combining all properties of FIRE with all properties of a TRUCK, or taking the intersection of the set of all things that are FIRE and the set of all things that are TRUCK, there is clearly more going on. An approach to deal with compositionality in conceptual spaces was suggested by Gärdenfors (2000a). In this approach, analogue to the English language, the order of the two concepts to be combined is important. Thus, when combining the concepts X and Y to XY, X acts as a modifier of Y, e.g. RED BRICK describes some kind of BRICK and BRICK RED describes some kind of RED. Effectively, the properties of X that correspond to properties of Y replace the latter. If Y does not have the properties of X, they are simply added; if Y does have overlapping properties with X, they are overruled[2]. An illustration of how compositionality can emergence from language game learning was reported by Vogt (2005). Common in this type of studies, the presence of learning mechanisms that can acquire compositional structures is assumed; as such the study investigates the conditions that favour the emergence of compositionality. Compositionality is implemented within a set of rewrite rules that refer to different aspects of the underlying conceptual space representation. It was shown experimentally that the presence of a 'transmission bottleneck' (in

[2]X may also influence certain other properties of Y. For instance, the addition of WOODEN to SPOON may yield a revision of the 'size' property of SPOON (making it bigger); or certain properties may be blocked as in STONE LION, where STONE blocks properties of LION like 'living' and 'sound' and 'habitat' as they do not apply to STONE. In summary, when combining X and Y, X may entail certain properties that can influence Y. Y however, determines the context in which these properties of X are applied.

which the teacher only uses part of it's language to teach a learning agent) favours the emergence of compositional structures within the shared language that agents develop.

Although considered fundamental to concepts, it was decided not to include theses topics in the model and simulations reported here. Yet, some effort was spent on investigating whether or not hierarchical structures could be included in the language game framework. In contrast to Vogt (2004), in which different hierarchical levels are modelled within the conceptual space representation of an agent's categories, the aim was to achieve hierarchical structures within the weights of the association matrix connecting word labels and categories, cf. O'Connor et al. (2009). It was found that in a basic implementation of language games there is no obvious way to facilitate different hierarchical levels of concepts by using the weights of the association matrix; the framework will have to be further extended as to support this. This endeavour was only partially finished and as such constitutes future work. A description of the exploration regarding if and how hierarchical structures might be incorporated in language game models through the association matrix is provided in appendix D.

8.3.2 Addition of associative networks and incorporation of LSA

As described in previous sections, the manner of modelling concepts in most of this work is based on a fairly straightforward representation of various quality dimensions in a conceptual space. Word labels are associated with regions in the conceptual space through social linguistic interacting which typically results in a one-to-one mapping between words and concepts. This one-to-one mapping allows for two-way activations, as perceiving a word activates associated concepts and the other way around, the perception of various quality data activates perceptual structures in the conceptual space and subsequently associated word labels. However, particularly in light of a typical property of the human memory, semantic priming[3] (Neely, 1977,

[3]Semantic priming refers to improvements in speed or accuracy to respond to a stimulus when this stimulus is preceded by another stimulus that is somehow semantically related.

1991), it seems prudent to accommodate some kind of semantic associative layer which places emphasis on linguistic association. More specifically, semantic priming can refer to priming effects that are purely semantic (e.g. 'dog'-'goat'; both words have similar meanings, they refer to furry mammal of comparable size) or to priming effects that are more associative in nature (e.g. 'bread'-'butter'; these words refer to concepts that are generally experienced in each others proximity, but they do not have a lot of overlapping perceptual features). It has been suggested that the human brain processes these two types of priming in a different manner (Chiarello et al., 1990). Furthermore, it has been shown that specifically priming effects based on pure semantic overlap (as opposed to associative priming effects) can be exhibited by models that extract a high dimensional semantic space from a co-occurrence matrix based on large text corpora (Lund et al., 1995). These findings illustrate that semantic priming, being a typical property of human memory and concept use, can be accounted for solely through analysis of statistical regularities in the linguistic data a system is exposed to.

As such, the model as described in chapter 3 could be augmented with mechanisms which form associations between words based on the statistical analysis of the linguistic context in which they are experienced. That is, it is envisioned that the additions of LSA-like techniques as described in section 2.4.3 might enable the system to perform more in line with human memory and concept use through the formation of an associative network-like structure which can facilitate semantic priming. In more practical terms, as the language game dynamics provide the formation of associations between words and concepts, incorporation of functionality as described above would entail the addition of a network layer which expresses associations between words only. This structure would be sensitive to statistical regularities in which words are experienced by the system and modify and update associative connections between words in the lexicon accordingly. Subsequently activation patterns of this linguistic network could influence lower level cognition and perception in a top-down fashion.

8.3.3 Further exploration of active and social learning in HRI

While the notion of active learning in machine learning has been fairly well established (Settles, 2009), active learning in combination with HRI is relatively novel. A key issue here is the question of how a robot learner should behave in order to effectively utilise active learning strategies while interacting with a human partner. The active learning strategies as reported in chapters 5 and 7 result in more effective robot learning. As such, they provide an illustration of how active learning strategies that have their basis in an algorithmic machine learning context can be extended to social human-robot interaction, thus broadening their application. While this highlights the importance of active and appropriate social participation of a robot learner, the strategies themselves are relatively simple. Also the learning task in which the robot and human teacher engaged is fairly straightforward and could be expanded, e.g. through a more challenging environment in which effective cooperation between robots and humans is required.

Indeed, more elaborated scenarios in which the robot employs more sophisticated means of interaction are imaginable. For instance, while the active component of the learning strategy employed in this work was based on a distance measurement to determine the least known concepts and as such implements only one active strategy, it is conceivable that a robot learner might have a repertoire of strategies to choose from. A combination of factors such as the state of the robot, the progress of the learning experience, the social context and potentially the preferred style of interaction with a human teacher can govern the selection of certain strategies over others.

Endowed with more social awareness, that is, a capacity to perceive and express a range of (non-verbal) social cues, a robot learner might be able to even more effectively modulate a learning experience. For instance, transparency on the robot side regarding its current understandings of a task and the deficits in its knowledge can help a human teacher to provide more effective instructions (Chao et al., 2010); appropriate turn-taking mechanisms can improve the interaction, increasing the

initiative of a human partner and resulting in improved task performance (Chao and Thomaz, 2012). Furthermore, as suggested by Cakmak et al. (2010), and also confirmed to certain degree through the observed gender differences in reception of an active learning robot (section 7.4.5.5), the optimal learning strategy may vary for different users. As such, robots will need to be receptive for these different user preferences.

In short, while the experiments reported in chapter 7 provide some insights in learning improvements that can be gained through social and active learning, it is expected that through increased social awareness and behaviour on the robot side learning through social interaction can improve even more.

Appendix A

Zoo dataset

Zoo dataset (UCI Machine Learning Repository; Frank and Asuncion, 2010)

- **MAMMAL (41)** aardvark, antelope, bear, boar, buffalo, calf, cavy, cheetah, deer, dolphin, elephant, fruitbat, giraffe, girl, goat, gorilla, hamster, hare, leopard, lion, lynx, mink, mole, mongoose, opossum, oryx, platypus, polecat, pony, porpoise, puma, pussycat, raccoon, reindeer, seal, sealion, squirrel, vampire, vole, wallaby, wolf
- **BIRD (20)** chicken, crow, dove, duck, flamingo, gull, hawk, kiwi, lark, ostrich, parakeet, penguin, pheasant, rhea, skimmer, skua, sparrow, swan, vulture, wren
- **REPTILE (5)** pitviper, seasnake, slowworm, tortoise, tuatara
- **FISH (13)** bass, carp, catfish, chub, dogfish, haddock, herring, pike, piranha, seahorse, sole, stingray, tuna
- **AMPHIBIAN (4)** frog, frog, newt, toad
- **INSECT (8)** flea, gnat, honeybee, housefly, ladybird, moth, termite, wasp
- **INVERTEBRATE (10)** clam, crab, crayfish, lobster, octopus, scorpion, seawasp, slug, starfish, worm

animal name	type	class	hair	feathers	eggs	milk	airborne	aquatic	predator	toothed	backbone	breathes	venomous	fins	legs	tail	domestic	catsize
aardvark	1	MAMMAL	1	0	0	1	0	0	1	1	1	1	0	0	4	0	0	1
antelope	1	MAMMAL	1	0	0	1	0	0	0	1	1	1	0	0	4	1	0	1
bear	1	MAMMAL	1	0	0	1	0	0	1	1	1	1	0	0	4	0	0	1
boar	1	MAMMAL	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	1
buffalo	1	MAMMAL	1	0	0	1	0	0	0	1	1	1	0	0	4	1	0	1
calf	1	MAMMAL	1	0	0	1	0	0	0	1	1	1	0	0	4	1	1	1
cavy	1	MAMMAL	1	0	0	1	0	0	0	1	1	1	0	0	4	0	1	0
cheetah	1	MAMMAL	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	1
deer	1	MAMMAL	1	0	0	1	0	0	0	1	1	1	0	0	4	1	0	1
dolphin	1	MAMMAL	0	0	0	1	0	1	1	1	1	1	0	1	0	1	0	1
elephant	1	MAMMAL	1	0	0	1	0	0	0	1	1	1	0	0	4	1	0	1
fruitbat	1	MAMMAL	1	0	0	1	1	0	0	1	1	1	0	0	2	1	0	0
giraffe	1	MAMMAL	1	0	0	1	0	0	0	1	1	1	0	0	4	1	0	1
girl	1	MAMMAL	1	0	0	1	0	0	1	1	1	1	0	0	2	0	1	1
goat	1	MAMMAL	1	0	0	1	0	0	0	1	1	1	0	0	4	1	1	1
gorilla	1	MAMMAL	1	0	0	1	0	0	0	1	1	1	0	0	2	0	0	1
hamster	1	MAMMAL	1	0	0	1	0	0	0	1	1	1	0	0	4	1	1	0
hare	1	MAMMAL	1	0	0	1	0	0	0	1	1	1	0	0	4	1	0	0
leopard	1	MAMMAL	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	1
lion	1	MAMMAL	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	1
lynx	1	MAMMAL	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	1
mink	1	MAMMAL	1	0	0	1	0	1	1	1	1	1	0	0	4	1	0	1
mole	1	MAMMAL	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	0
mongoose	1	MAMMAL	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	1
opossum	1	MAMMAL	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	0
oryx	1	MAMMAL	1	0	0	1	0	0	0	1	1	1	0	0	4	1	0	1
platypus	1	MAMMAL	1	0	1	1	0	1	1	0	1	1	0	0	4	1	0	1
polecat	1	MAMMAL	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	1
pony	1	MAMMAL	1	0	0	1	0	0	0	1	1	1	0	0	4	1	1	1
porpoise	1	MAMMAL	0	0	0	1	0	1	1	1	1	1	0	1	0	1	0	1
puma	1	MAMMAL	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	1
pussycat	1	MAMMAL	1	0	0	1	0	0	1	1	1	1	0	0	4	1	1	1
raccoon	1	MAMMAL	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	1
reindeer	1	MAMMAL	1	0	0	1	0	0	0	1	1	1	0	0	4	1	1	1
seal	1	MAMMAL	1	0	0	1	0	1	1	1	1	1	0	1	0	0	0	1
sealion	1	MAMMAL	1	0	0	1	0	1	1	1	1	1	0	1	2	1	0	1
squirrel	1	MAMMAL	1	0	0	1	0	0	0	1	1	1	0	0	2	1	0	0
vampire	1	MAMMAL	1	0	0	1	1	0	0	1	1	1	0	0	2	1	0	0
vole	1	MAMMAL	1	0	0	1	0	0	0	1	1	1	0	0	4	1	0	0
wallaby	1	MAMMAL	1	0	0	1	0	0	0	1	1	1	0	0	2	1	0	1
wolf	1	MAMMAL	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	1
chicken	2	BIRD	0	1	1	0	1	0	0	0	1	1	0	0	2	1	1	0
crow	2	BIRD	0	1	1	0	1	0	1	0	1	1	0	0	2	1	0	0
dove	2	BIRD	0	1	1	0	1	0	0	0	1	1	0	0	2	1	1	0
duck	2	BIRD	0	1	1	0	1	1	0	0	1	1	0	0	2	1	0	0
flamingo	2	BIRD	0	1	1	0	1	0	0	0	1	1	0	0	2	1	0	1
gull	2	BIRD	0	1	1	0	1	1	1	0	1	1	0	0	2	1	0	0
hawk	2	BIRD	0	1	1	0	1	0	1	0	1	1	0	0	2	1	0	0
kiwi	2	BIRD	0	1	1	0	0	0	1	0	1	1	0	0	2	1	0	0
lark	2	BIRD	0	1	1	0	1	0	0	0	1	1	0	0	2	1	0	0
ostrich	2	BIRD	0	1	1	0	0	0	0	0	1	1	0	0	2	1	0	1
parakeet	2	BIRD	0	1	1	0	1	0	0	0	1	1	0	0	2	1	1	0
penguin	2	BIRD	0	1	1	0	0	1	1	0	1	1	0	0	2	1	0	1
pheasant	2	BIRD	0	1	1	0	1	0	0	0	1	1	0	0	2	1	0	0
rhea	2	BIRD	0	1	1	0	0	0	1	0	1	1	0	0	2	1	0	1
skimmer	2	BIRD	0	1	1	0	1	1	1	0	1	1	0	0	2	1	0	0
skua	2	BIRD	0	1	1	0	1	1	1	0	1	1	0	0	2	1	0	0
sparrow	2	BIRD	0	1	1	0	1	0	0	0	1	1	0	0	2	1	0	0
swan	2	BIRD	0	1	1	0	1	1	0	0	1	1	0	0	2	1	0	1
vulture	2	BIRD	0	1	1	0	1	0	1	0	1	1	0	0	2	1	0	1
wren	2	BIRD	0	1	1	0	1	0	0	0	1	1	0	0	2	1	0	0

Table A.1: Full Zoo dataset displaying all properties.

animal name	type	class	hair	feathers	eggs	milk	airborne	aquatic	predator	toothed	backbone	breathes	venomous	fins	legs	tail	domestic	catsize
pitviper	3	REPTILE	0	0	1	0	0	0	1	1	1	1	1	0	0	1	0	0
seasnake	3	REPTILE	0	0	0	0	0	1	1	1	1	0	1	0	0	1	0	0
slowworm	3	REPTILE	0	0	1	0	0	0	1	1	1	1	0	0	0	1	0	0
tortoise	3	REPTILE	0	0	1	0	0	0	0	0	1	1	0	0	4	1	0	1
tuatara	3	REPTILE	0	0	1	0	0	0	1	1	1	1	0	0	4	1	0	0
bass	4	FISH	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	0
carp	4	FISH	0	0	1	0	0	1	0	1	1	0	0	1	0	1	1	0
catfish	4	FISH	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	0
chub	4	FISH	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	0
dogfish	4	FISH	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	1
haddock	4	FISH	0	0	1	0	0	1	0	1	1	0	0	1	0	1	0	0
herring	4	FISH	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	0
pike	4	FISH	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	1
piranha	4	FISH	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	0
seahorse	4	FISH	0	0	1	0	0	1	0	1	1	0	0	1	0	1	0	0
sole	4	FISH	0	0	1	0	0	1	0	1	1	0	0	1	0	1	0	0
stingray	4	FISH	0	0	1	0	0	1	1	1	1	0	1	1	0	1	0	1
tuna	4	FISH	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	1
frog	5	AMPHIBIAN	0	0	1	0	0	1	1	1	1	1	0	0	4	0	0	0
frog	5	AMPHIBIAN	0	0	1	0	0	1	1	1	1	1	1	0	4	0	0	0
newt	5	AMPHIBIAN	0	0	1	0	0	1	1	1	1	1	0	0	4	1	0	0
toad	5	AMPHIBIAN	0	0	1	0	0	1	0	1	1	1	0	0	4	0	0	0
flea	6	INSECT	0	0	1	0	0	0	0	0	0	1	0	0	6	0	0	0
gnat	6	INSECT	0	0	1	0	1	0	0	0	0	1	0	0	6	0	0	0
honeybee	6	INSECT	1	0	1	0	1	0	0	0	0	1	1	0	6	0	1	0
housefly	6	INSECT	1	0	1	0	1	0	0	0	0	1	0	0	6	0	0	0
ladybird	6	INSECT	0	0	1	0	1	0	1	0	0	1	0	0	6	0	0	0
moth	6	INSECT	1	0	1	0	1	0	0	0	0	1	0	0	6	0	0	0
termite	6	INSECT	0	0	1	0	0	0	0	0	0	1	0	0	6	0	0	0
wasp	6	INSECT	1	0	1	0	1	0	0	0	0	1	1	0	6	0	0	0
clam	7	INVERTEBRATE	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0
crab	7	INVERTEBRATE	0	0	1	0	0	1	1	0	0	0	0	0	4	0	0	0
crayfish	7	INVERTEBRATE	0	0	1	0	0	1	1	0	0	0	0	0	6	0	0	0
lobster	7	INVERTEBRATE	0	0	1	0	0	1	1	0	0	0	0	0	6	0	0	0
octopus	7	INVERTEBRATE	0	0	1	0	0	1	1	0	0	0	0	0	8	0	0	1
scorpion	7	INVERTEBRATE	0	0	0	0	0	0	1	0	0	1	1	0	8	1	0	0
seawasp	7	INVERTEBRATE	0	0	1	0	0	1	1	0	0	0	1	0	0	0	0	0
slug	7	INVERTEBRATE	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0

Table A.2: Full Zoo dataset displaying all properties continued

Appendix B

Colour conversion models

To model varying responsiveness of cone receptors, we first convert RGB colour triplets to CIE XYZ colour space and then convert to LMS colour space (Fairchild, 2005). Next we weigh the LMS values, as to model the varying responsiveness of the L, M and S cones.

The LMS colour space gives the response of the three types of cones. For the conversion from RGB to CIE XYZ we first need to linearise RGB values, this is done by applying an inverse sRGB companding function to the R, G and B values

$$v = \begin{cases} V/12.92 & V \leq 0.04045 \\ \left(\frac{V+0.055}{1.055}\right)^{2.4} & V > 0.04045 \end{cases}$$

next the linearized values r, g and b are converted to CIE XYZ using

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = M_{sRGB \to XYZ} \begin{bmatrix} r \\ g \\ b \end{bmatrix}$$

with $M_{sRGB \to XYZ}$ being the sRGB conversion matrix, with white reference D65

$$M_{sRGB \to XYZ} = \begin{bmatrix} 0.4124564 & 0.3575761 & 0.1804375 \\ 0.2126729 & 0.7151522 & 0.0721750 \\ 0.0193339 & 0.1191920 & 0.9503041 \end{bmatrix}.$$

The conversion from CIE XYZ to LMS is given by:

$$\begin{bmatrix} L \\ M \\ S \end{bmatrix} = M_{XYZ \to LMS} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$$

with

$$M_{XYZ \to LMS} = \begin{bmatrix} 0.8951 & 0.2664 & -0.1614 \\ -0.7502 & 1.7135 & 0.0367 \\ 0.0389 & -0.0685 & 1.0296 \end{bmatrix}$$

To model the varying responsiveness of the cones, each agent a weighs the L and M response with a weight α_a and β_a, randomly drawn from the range $[0.33, 1]$.

$$\begin{bmatrix} L' \\ M' \\ S' \end{bmatrix} = \begin{bmatrix} \alpha_a L \\ \beta_a M \\ S \end{bmatrix}.$$

Appendix C

Social learning experiment data

“I would like to learn this one”
“could you teach me this one?”
“this one looks interesting”
“now, what about this one?”
“this is interesting”
“em, what about this one?”
“what about this one?”
“I would like to know what this is”
“ok, what do we have here?”
“yes, this looks interesting”
“what about this one?”
“em, I would like to know what this is”

Table C.1: Statements made by the active learning robot.

participant #	AL	tot-case	conf-case	%
1	0	15	4	26.67%
2	0	12	11	91.67%
3	1	16	11	68.75%
4	1	15	10	66.67%
5	0	15	6	40.00%
8	1	19	10	52.63%
9	1	18	6	33.33%
10	1	14	10	71.43%
11	0	10	2	20.00%
12	1	18	10	55.56%
13	0	18	3	16.67%
14	0	19	9	47.37%
15	0	10	8	80.00%
16	1	17	8	47.06%
17	0	14	4	28.57%
18	0	14	8	57.14%
19	1	17	10	58.82%
20	1	13	7	53.85%
21	1	10	6	60.00%
22	0	20	7	35.00%
23	1	15	6	40.00%
24	0	18	5	27.78%
25	0	23	6	26.09%
26	1	15	9	60.00%
27	1	15	10	66.67%
28	1	15	10	66.67%
29	0	16	6	37.50%
30	0	17	9	52.94%
31	0	15	8	53.33%
32	1	7	4	57.14%
33	0	21	8	38.10%
34	1	14	9	64.29%
35	1	17	7	41.18%
36	0	20	8	40.00%
37	1	15	8	53.33%
38	0	13	4	30.77%
39	1	13	7	53.85%
40	0	16	12	75.00%
41	0	18	10	55.56%

Table C.2: Ambiguity interpretation for all participants. AL denotes the AL (1) or non-AL (0) condition, 'tot-case' denotes the number of cases with two exemplars from the same category, 'conf-case' denotes the number of cases in which participants confirmed the robot's guess as being correct and % calculates the percentage ('conf-case' divided by 'tot-case').

participant #	AL	GG success	AL response
1	no	0.48	0.3
2	no	0.64	0.28
5	no	0.56	0.38
11	no	0.6	0.3
13	no	0.46	0.26
14	no	0.54	0.44
15	no	0.68	0.32
17	no	0.56	0.38
18	no	0.64	0.22
22	no	0.54	0.22
24	no	0.4	0.38
25	no	0.42	0.38
29	no	0.62	0.28
30	no	0.68	0.48
31	no	0.54	0.3
33	no	0.48	0.36
36	no	0.62	0.26
38	no	0.54	0.32
40	no	0.7	0.3
41	no	0.62	0.18
3	yes	0.72	0.48
4	yes	0.66	0.76
8	yes	0.62	0.94
9	yes	0.52	0.4
10	yes	0.68	0.46
12	yes	0.64	0.58
16	yes	0.46	0.38
19	yes	0.6	0.46
20	yes	0.62	0.62
21	yes	0.76	0.44
23	yes	0.54	0.46
26	yes	0.62	0.8
27	yes	0.56	0.44
28	yes	0.7	0.44
32	yes	0.7	0.58
34	yes	0.72	0.42
35	yes	0.56	0.38
37	yes	0.62	0.86
39	yes	0.6	0.8

Table C.3: Guessing game success and response to AL of individual participants for the AL and non-AL group.

C.1 Questionnaire

Social Robot Teaching Questionnaire

Participant number:	**Age:**
Gender: F / M	**Native speaker: yes / no**

Please answer the following questions by placing an 'X' on the spot that best reflects your answer. Additionally, you can provide comments to elaborate your answers.

1. How do you rate your interaction with the robot?

not satisfactory at all **very satisfactory**

comments

2. How do you rate the robot's behaviour?

not natural at all **very natural**

comments

3. Do you have any experience with robots?

I have no experience with robots **I have a lot of experience with robots**

comments

4. Who was in control of the teaching sessions?

I was in control **the robot was in control**

comments

5. On what basis did you choose the animal examples as topic? Please explain.

6. Do you like science fiction (books, film, etc)?

I don't like science fiction at all **I very much like science fiction**

comments

7. How many emotions do you think the robot has?

the robot has no emotions **the robot has a lot of emotions**

comments

8. How smart do you think the robot is?

the robot is not smart at all **the robot is very smart**

comments

9. How many hours per week do you spend using a computer?

hours computer use per week (estimate):

comments

10. General comments

How I am in general

Here are a number of characteristics that may or may not apply to you. For example, do you agree that you are someone who *likes to spend time with others*? Please write a number next to each statement to indicate the extent to which **you agree or disagree with that statement.**

1 Disagree Strongly	**2** Disagree a little	**3** Neither agree nor disagree	**4** Agree a little	**5** Agree strongly

I am someone who…

1. _____ Is talkative
2. _____ Tends to find fault with others
3. _____ Does a thorough job
4. _____ Is depressed, blue
5. _____ Is original, comes up with new ideas
6. _____ Is reserved
7. _____ Is helpful and unselfish with others
8. _____ Can be somewhat careless
9. _____ Is relaxed, handles stress well
10. _____ Is curious about many different things
11. _____ Is full of energy
12. _____ Starts quarrels with others
13. _____ Is a reliable worker
14. _____ Can be tense
15. _____ Is ingenious, a deep thinker
16. _____ Generates a lot of enthusiasm
17. _____ Has a forgiving nature
18. _____ Tends to be disorganized
19. _____ Worries a lot
20. _____ Has an active imagination
21. _____ Tends to be quiet
22. _____ Is generally trusting
23. _____ Tends to be lazy
24. _____ Is emotionally stable, not easily upset
25. _____ Is inventive
26. _____ Has an assertive personality
27. _____ Can be cold and aloof
28. _____ Perseveres until the task is finished
29. _____ Can be moody
30. _____ Values artistic, aesthetic experiences
31. _____ Is sometimes shy, inhibited
32. _____ Is considerate and kind to almost everyone
33. _____ Does things efficiently
34. _____ Remains calm in tense situations
35. _____ Prefers work that is routine
36. _____ Is outgoing, sociable
37. _____ Is sometimes rude to others
38. _____ Makes plans and follows through with them
39. _____ Gets nervous easily
40. _____ Likes to reflect, play with ideas
41. _____ Has few artistic interests
42. _____ Likes to cooperate with others
43. _____ Is easily distracted
44. _____ Is sophisticated in art, music, or literature

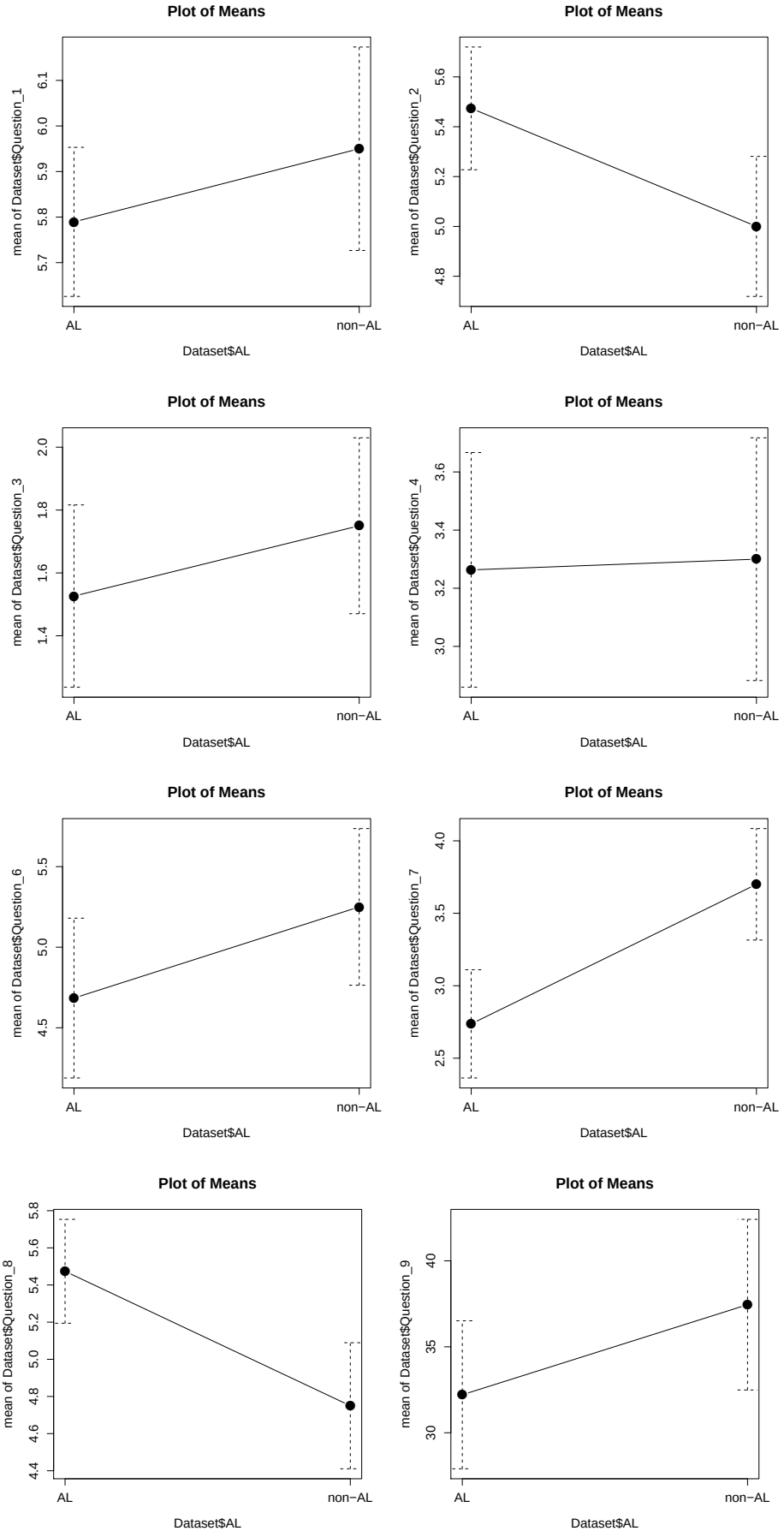


Figure C.1: Questionnaire response based on AL and non-AL groups

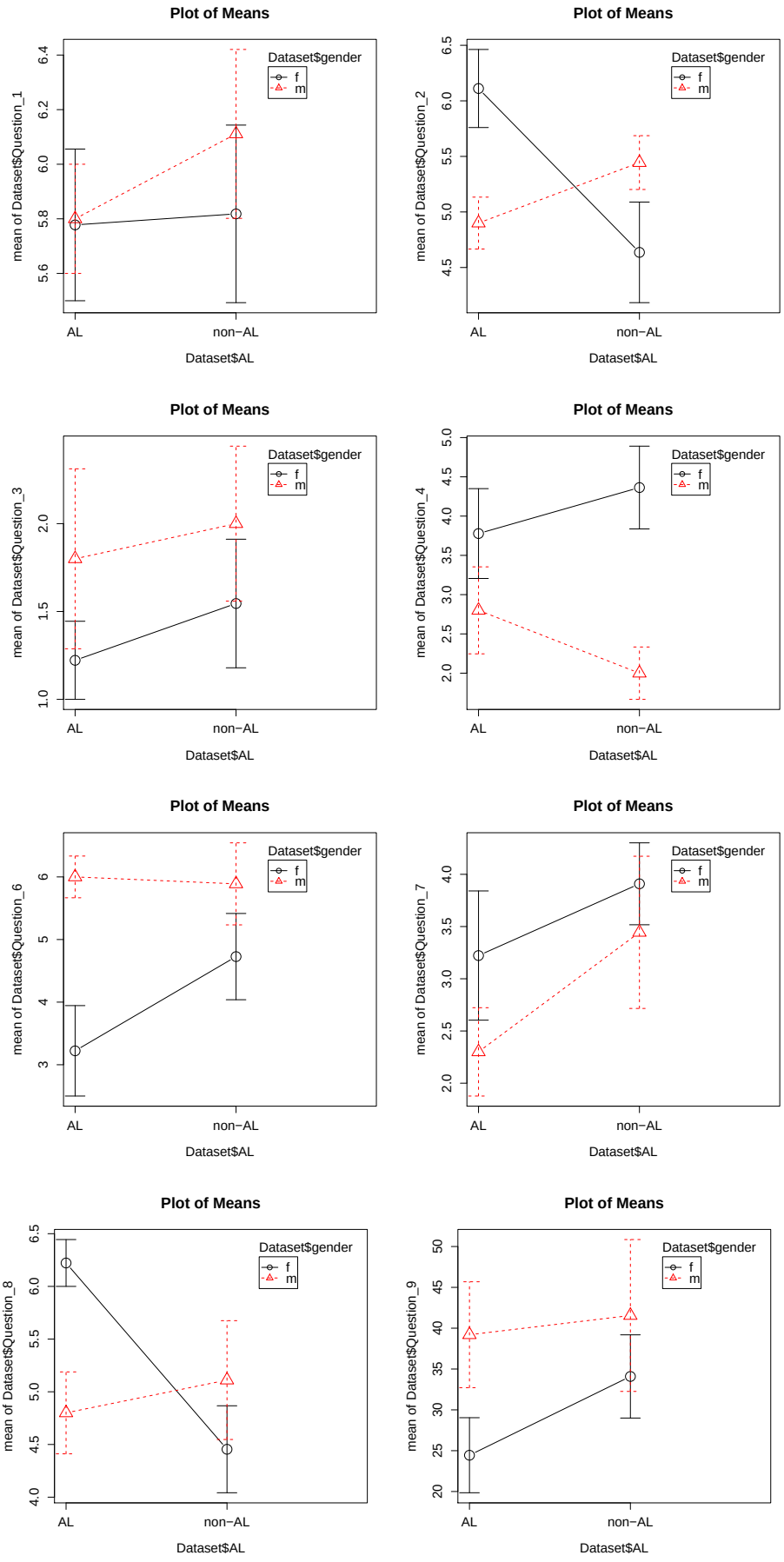


Figure C.2: Questionnaire response split in gender and AL and non-AL groups

C.1.1 Participants' response to Question 5: "On what basis did you choose the animal examples as topic?"

- "Based on reproductive system"
- "what I thought would be easy for the robot to learn"
- "often where the robot looked to, what seemed to be the easiest for me"
- "1. I choose this animal which I was familiar 2. I choose this picture where the robot was looking to."
- "As randomly as possible, though I attempted to introduce new animals when I had the opportunity"
- "random"
- "By categories I was sure on. Ignored ones I was not sure even I knew."
- "The picture I felt may challenge the robot"
- "Ones I had not taught the robot. Chose different examples of similar animals to teach the robot difference in types."
- "Something that had not appeared for a while, or trickier options such as a chicken being a bird, but looking like a mammal."
- "on preference, e.g. because I like a whale better than a bird and based on what animal I was sure of knowing the right category for"
- "first thing that came to sight."
- "Ones I like"
- "from my knowledge, learning in school. Also watching nature programs"
- "I wanted to show him the widest range of categories, but also to check he was remembering what he had learned."
- "firstly the pictures I liked the most or the animals that were my favourite. Then tried to include the rest when they came up"
- "I tended to choose an animal I prefered."
- "On what I was sure of or what the robot stated & looked at when saying 'I want to learn this one'."
- "I choose a variety to ensure learning."
- "At random if I knew all three, if not I would select the one I knew. If two were the same I would pick one of them."
- "When 2 animals were same category I chose 3rd animal easier for the robot to learn. When 2+ animals were the same, I made it make a mistake to learn new animals."
- "whichever one appealed to me, either cuteness or weirdness"
- "when new animals not encountered would pick or animals of group covered but not this individual species"

- "I've tried to repeat the same groups a couple of times. I avoided one of the groups as I wasn't sure what it means :)"
- "I tried to choose ones the robot would learn from."
- "I had to be certain about the category (i.e., not sure? Whether certain animals were birds or mammals → platypus)"
- "the categories I actually knew! (have to brush up my biology!!) and the attractiveness of the pictures"
- "The ones I was certain I knew the categories"
- "what the pictures looked like. Animals I am familiar of"
- "which animal I prefered, how cute & fluffy it looked, or how interesting I thought it was."
- "The animals I was sure about their category."
- "I chose ones that I though would be more difficult to guess to see how good the program was"
- "- random at first then tried to pick some I know I had taught (ie -used same photo) and then test with other of same category"
- "which ever one stood out the most"
- "I chose based on differences to other parts of the same category such as seals, dolphins & big cats or based on similarity to others & insects"
- "sometimes the one the robot looked at. Sometimes when there was just one separate example among two similar, sometimes when there was a typical example of a category"
- "I chose the easiest, those one are easily to be classified"
- "what I recognised first or if one looked more interesting"

Appendix D

Modelling hierarchical concepts in language games

In human cognitive systems concepts tend to have a hierarchical structure. That is, concepts do not exist in isolation, but are embedded within a taxonomy of related other concepts. Typically, this taxonomy is depicted as a hierarchical tree structure, with the most general (superordinate) concepts at the top, the 'ordinary' concepts in the middle (basic level) and the most specific concepts at the bottom (subordinate). In such a taxonomy, items in the tree are instantiations of their parent items or inhered properties from them, thus defining the hierarchy. So, a golden retriever *isa* dog *isa* animal *isa* living *isa* object. There is abundant evidence that 'basic level' concepts have special status (Rosch, 1978), i.e. are the most accessible, the most used etc. Also in child-directed speech caregivers typically use the basic level to describe something to a child.

It is not clear to what extent people's use of concepts is truly organised in strict taxonomic fashion; i.e. some natural taxonomies like animal order are used be people, but it could be argued that this is learned trough acquisition of knowledge about genetic inheritance, rather than a 'real' feat of conceptual organisation. CAT and DOG are both ANIMAL is a very typical example of this, but all it would require is to learn that ANIMAL has strong connections to the perceptual properties of both CAT and DOG. O'Connor et al. (2009) show that this kind of hierarchical organisation can spontaneously arise by statistical learning of properties of objects.

O'Connor et al. show how a 2-level hierarchy (basic level and superordinate) can be represented in a feature-based attractor network. After training the network is able to express basic level concepts with very specific features (high weight for a few defining features) and superordinates with a more moderate pattern of feature activation.

D.1 Hierarchical CS

Hierarchy could be strictly imposed in a model, i.e. every concept is placed at a certain level on a taxonomy. However, 'functional' hierarchy could also be achieved by having a network of relations in which a hierarchy can be build through hierarchical clustering algorithm (based on distance).

Based on the work of O'Connor et al. (2009), we view an agent as having knowledge on different hierarchical levels when it knows some words with a strong connection to only a few features, and some other words with relative weak connections to a wide range of features. The former words would resemble something at the level of basic level words, and the latter would be more general, at the level of superordinates. Using this as a staring point, we can endow an agent with some predefined knowledge at different hierarchical levels. Such a typical representation is displayed in figure D.1. In here the agent has a repertoire of words to indicate colour names and basic shapes that have a strong connection to one particular point in the agents CS. Next to this the agent knows the words 'colour' and 'shape' which have a weaker connection (0.5) to a range of points in the CS. The latter can be thought of as superordinate concepts.

However, when an agent endowed with the conceptual knowledge as described above acts as a teacher to another agent with a blank repertoire while engaging in standard language game interaction, the learning agent fails to learn the superordinates. This is the case because the due to the weak link of the superordinate words, they are never used to express an item in the context, making it impossible for the learning agent to acquire these words.

In a natural situation superordinate words would be used (i.e. by caregivers) in

	kaeluq	wkfp7z	iztprp	0l7us7	0tb8at	ls3qs8	z7mn1w	bg36t5	1zzund	gz1rg9	u3u1u8	v35zo1	pn6n63	tsjxh1	24waib
VALUES	[1. 0. 0.]	[1. 0. 1.]	[0.5 0. 1	[0. 0. 1.]	[0.300000	[0.600000	[0. 0. 0.]	[0. 1. 0.]	[1. 1. 1.]	[0.5 0.5	[1. 1. 0.]	[10.]	[3.]	[4.]	[8.]
red	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
pink	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
purple	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
blue	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
turquoise	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
brown	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
black	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
green	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
white	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
grey	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
yellow	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
circle	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
triangle	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
square	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
pentagon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
colour	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0	0	0	0
shape	0	0	0	0	0	0	0	0	0	0	0	0.5	0.5	0.5	0.5

Figure D.1: Agent knowledge

cases when they don't just want to name a certain object, but rather speak of the class of objects that is represented by a particular superordinate. For instance "lets find all the animals", when reading a book to a child. Another situation would be if it is required to contrast a certain object against a group of other type of objects. I.e. "where is the animal" when trying to discriminate a (picture) of an dog from pictures of other (non-living) objects. Describing the dog in terms of its property of being animate contrasts it to the other objects.

In O'Connor et al. (2009) the learning network is first exposed to basic level data and after that it given superordinate training data. However, rather than doing this manual, we seek to create a situation in which the speaking agent has good reason to use the superordinate word as a description of the topic, instead of a more specific label.

Ideally, within LG the topic should be described occasionally on a superordinate level in order to have the learning agent acquire superordinate labels. One way of doing this is to take other items of the context into account. By doing this, a label that is the most descriptive in terms of the overall context can be used. For instance, if there are 3 items in the context, of which the topic is superordinate A, and the two others are superordinate B, than it should be more descriptive to use A to label the topic, rather than a more specific C, D, E or F basic level labels.

Another way of having an agent occasionally using the superordinate label is by modifying the way in which labels are selected to describe a stimulus. Normally, the label with the strongest connection to an observed stimulus would be used,

but this can be modified into a probabilistic version in which the strength of the connection determines the chances of selecting the label. So, if an agent is endowed with predefined knowledge, with labels connecting strongly to specific (basic level) concepts and somewhat weaker to superordinates, the probability of using the basic level label would be high and of using the superordinate label would be lower, but nevertheless present.

Why basic LG modelling is not rich enough for hierarchy representations. The assumption that hierarchies can be represented in a weighted matrix as illustrated in Figure D.1 is based on the observations of human hierarchical naming and modelling using a network architecture, as done by O'Connor et al. (2009). Using this kind of representation, a 'basic level' word would be characterised by having relatively few but strong connections to a limited number of perceptual features, whereas a 'superordinate' word would have a relatively large number of weaker connections to perceptual features.

To model this a teaching agent was equipped with the relevant hierarchical structure as described above, and engaged in interaction with a learning agent. To make sure that not only words with the highest connection to perceptions were used (as would be typical in a normal LG), the choice of word labels was modified in such a way that the strength of the connection specifies a chance of using that word. More specifically, all the weights were normalised to ensure a total weight of 1.0, so that having a weight of 0.4 encodes for 40% chance of using that word label. Thus, during LG interaction, a teaching agent would most often use the basic level words (because they have the strongest connection), but would occasionally also employ superordinate words. Running a LG like this results in the learning more or less learning the basic level words, and thus achieving some kind of performance which is above chance level. However, the learning of superordinates is problematic, as typically the superordinate also tends to become strongly associated with one specific percept, rather than a weak connection to multiple percepts. This is due to the nature of the language game. The specific percept that gets associated with the superordinate label tends to be a percept with a central position in the input

space, i.e. a greyish colour when the colour domain is used, as (0.5, 0.5, 0.5) is most general with respect to all possible colours in RGB encoding. However, rather than concluding that this percept represents a superordinate, it also encodes for the basic level of grey.

The inability to represent hierarchical structures in LG representation is due to the manner of organising associations in the word label - percept matrix. Having a weak connection does not really specify that a certain word has a certain connection to a percept, but rather encodes that it is not really effective in LG interaction. This is related to the 'winner takes all' strategy that is employed, a language games aim to find single effective connections between words and percepts, rather than associating one word with multiple percepts. Synonymy is avoided through the mechanism of lateral inhibition.

The manner in which training data was generated was also modified. Rather than having pure random points in the input space (as in basic LG), training data was generated with the use of prototypes. Thus, each item in the context would be based on a prototypical structure in such a way the token generated from the (proto)type would always be closer to its own type than to any other of the prototypes.

The communicative success of a language game with such a probabilistic selection mechanism is shown in figure D.1, and the teaching agent structure is depicted in figure D.3, before and after learning.

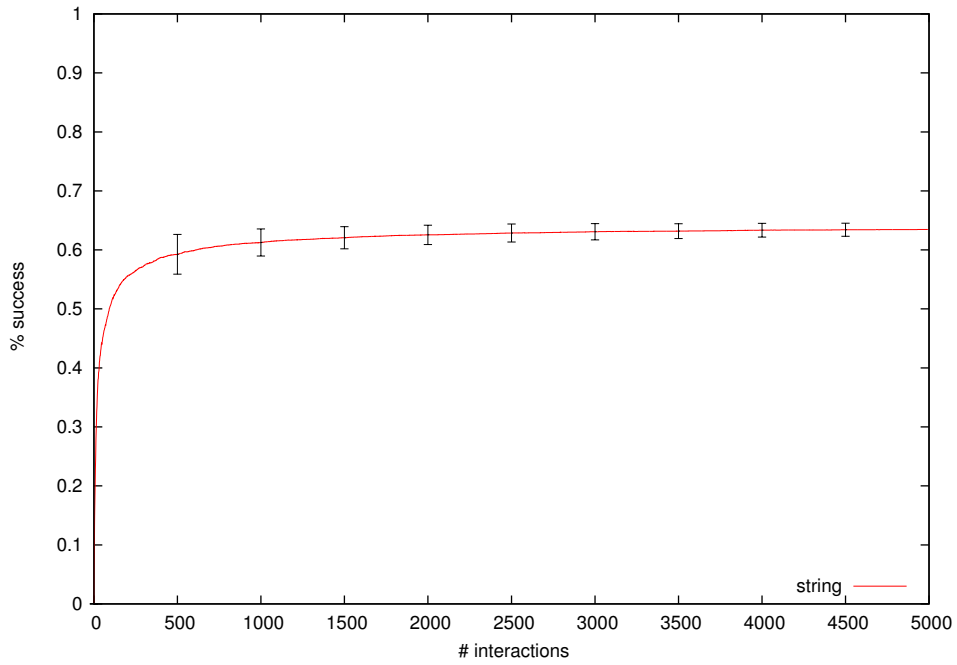


Figure D.2: LG with probabilistic label selection, 5000 interactions, 100 replicas

	dlc7qb	if6uy6	33vwbm	s0f1zh
VALUES	[1. 2. 4. 3.]	[2. 2. 4. 1.]	[3. 1. 4. 3.]	[4. 3. 2. 1.]
C	1	0	0	0
D	0	1	0	0
A	0.5	0.5	0	0
E	0	0	1	0
F	0	0	0	1
B	0	0	0.5	0.5

	ai3asl	6meffw	8l10pv	26k1qa
VALUES	[1. 2. 4. 3.]	[2. 2. 4. 1.]	[3. 1. 4. 3.]	[4. 3. 2. 1.]
C	1	0	0	0
D	0	0.99	0	0
A	0.01	0.99	0	0
E	0	0	0.99	0
F	0	0	0	1
B	0	0	0.99	0.01

Figure D.3: Teaching agent knowledge, before and after

Bibliography

Adams, B. and Raubal, M. (2009). A metric conceptual space algebra. In *Proceedings of the 9th international conference on Spatial information theory*, pages 51–68. Springer-Verlag.

Aisbett, J. and Gibbon, G. (2001). A general formulation of conceptual spaces as a meso level representation. *Artificial Intelligence*, 133(12):189 – 232.

Akhtar, N. and Tomasello, M. (2000). The social nature of words and word learning. *Becoming a word learner: A debate on lexical acquisition*, pages 115–135.

Al Moubayed, S., Edlund, J., and Beskow, J. (2012). Taming Mona Lisa: communicating gaze faithfully in 2D and 3D facial projections. *ACM Transactions on Interactive Intelligent Systems*, 1(2):25.

Allen, C. and Hauser, M. (1991). Concept attribution in nonhuman animals: Theoretical and methodological problems in ascribing complex mental processes. *Philosophy of Science*, pages 221–240.

Anderson, M. (2003). Embodied cognition: A field guide. *Artificial intelligence*, 149(1):91–130.

Asada, M., Hosoda, K., Kuniyoshi, Y., Ishiguro, H., Inui, T., Yoshikawa, Y., Ogino, M., and Yoshida, C. (2009). Cognitive developmental robotics: A survey. *Autonomous Mental Development, IEEE Transactions on*, 1(1):12–34.

Asada, M., MacDorman, K., Ishiguro, H., and Kuniyoshi, Y. (2001). Cognitive developmental robotics as a new paradigm for the design of humanoid robots. *Robotics and Autonomous Systems*, 37(2):185–193.

Atienza, R. and Zelinsky, A. (2002). Active gaze tracking for human-robot interaction. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, ICMI '02, pages 261–, Washington, DC, USA. IEEE Computer Society.

Baldwin, D. (1995). Understanding the link between joint attention and language. In Moore, C. and Dunham, P., editors, *Joint attention Its origins and role in development*, pages 131–158. Lawrence Erlbaum, Hillsdale, NJ.

Baldwin, D. and Moses, L. (2001). Links between social understanding and early word learning: Challenges to current accounts. *Social Development*, 10:309–329.

Baronchelli, A., Gong, T., Puglisi, A., and Loreto, V. (2010). Modeling the emergence of universality in color naming patterns. *Proceedings of the National Academy of Sciences*, 107(6):2403–2407.

Barsalou, L. (1982). Context-independent and context-dependent information in concepts. *Memory & Cognition*, 10:82–93. 10.3758/BF03197629.

Barsalou, L. (1999). Perceptual symbol systems. *Behavioral and brain sciences*, 22(04):577–660.

Barsalou, L. (2008). Grounded cognition. *Annu. Rev. Psychol.*, 59:617–645.

Bartneck, C., Kulić, D., Croft, E., and Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1(1):71–81.

Baxter, P., de Greeff, J., Wood, R., and Belpaeme, T. (2012). "And what is a seasnake?" Modelling the acquisition of concept prototypes in a developmental framework. In *Proceedings of the joint International Conference on Developmental Learning (ICDL) & Epigenetic Robotics 2012*, San Diego, USA. IEEE.

Baxter, P., Wood, R., Morse, A., and Belpaeme, T. (2011). Memory-Centred Architectures: Perspectives on Human-level Cognitive. In *Proceedings of the Advances in Cognitive Systems tracks at the AAAI Fall Symposium 2011*, pages 26–33. AAAI Press.

Belpaeme, T. (2002a). Communicating colour embedded in a colour context: Report on experiments on communicating colour meaning between Flemish informants. Technical Report 2002-10, Artificial Intelligence Lab, Vrije Universiteit Brussel.

Belpaeme, T. (2002b). *Factors influencing the origins of colour categories*. PhD thesis, Vrije Universiteit Brussel, Artificial Intelligence Lab.

Belpaeme, T., Baxter, P., Read, R., Wood, R., Cuayáhuitl, H., Kiefer, B., Racioppa, S., Kruijff-Korbayová, I., Athanasopoulos, G., Enescu, V., Looije, R., Neerincx, M., Demiris, Y., Ros-Espinoza, R., Beck, A., Cañamero, L., Hiolle, A., Lewis, M., Baroni, I., Nalin, M., Cosi, P., Paci, G., Tesser, F., Sommavilla, G., and Humbert, R. (2013). Multimodal Child-Robot Interaction: Building Social Bonds. *Journal of Human-Robot Interacion*, 1(2):33–53.

Belpaeme, T. and Bleys, J. (2005). Explaining universal colour categories through a constrained acquisition process. *Adaptive Behavior*, 13(4):293–310.

Belpaeme, T. and Morse, A. (2010). Time will tell - why it is too early to worry. *Interaction studies*, 11(2):191–195.

Belpaeme, T. and Morse, A. (2012). Word and category learning in a continuous semantic domain: Comparing cross-situational and interactive learning. *Advances in Complex Systems*, 15(03n04).

de Beule, J. and Bleys, J. (2010). Self-organization and emergence in language a case study for color. In Smith, A. D. M., Schouwstra, M., de Boer, B., and Smith, K., editors, *Proceedings of the 8th International Conference on the Evolution of Language*, pages 83–90. World Scientific.

Bickerton, D. (1995). *Language and human behavior*. University of Washington Press.

Biederman, I. (1985). Human image understanding: Recent research and a theory. *Computer vision, Graphics, and image processing*, 32(1):29–73.

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

Boroditsky, L. (2006). Linguistic relativity. In *Encyclopedia of Cognitive Science*. John Wiley & Sons.

Brainard, D. H., Roorda, A., Yamauchi, Y., Calderone, J. B., Metha, A., Neitz, M., Neitz, J., Williams, D. R., and Jacobs, G. H. (2000). Functional consequences of the relative numbers of l and m cones. *J. Opt. Soc. Am. A*, 17(3):607–614.

Braitenberg, V. (1986). *Vehicles: Experiments in synthetic psychology*. The MIT press.

Breazeal, C. (2000). *Sociable machines: Expressive social exchange between humans and robots*. PhD thesis, MIT.

Breazeal, C. (2004). Social interactions in hri: the robot view. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 34(2):181 –186.

Breazeal, C., Brooks, A., Gray, J., Hoffman, G., Kidd, C., Lee, H., Lieberman, J., Lockerd, A., and Mulanda, D. (2004). Humanoid robots as cooperative partners for people. *Int. Journal of Humanoid Robots*, 1(2):1–34.

Breazeal, C., Kidd, C. D., Thomaz, A. L., Hoffman, G., and Berlin, M. (2005). Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*, pages 708–713. IEEE.

Breazeal, C. and Scassellati, B. (2000). Infant-like social interactions between a robot and a human caregiver. *Adaptive Behavior*, 8(1):49–74.

Brooks, A., Gray, J., Hoffman, G., Lockerd, A., Lee, H., and Breazeal, C. (2004). Robot's play: interactive games with sociable machines. *Computers in Entertainment (CIE)*, 2(3):10–10.

Brooks, R. A. (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, 2(1):14–23.

Burton, A. M. (1994). Learning new faces in an interactive activation and competition model. *Visual Cognition*, 1(2-3):313–348.

Cakmak, M., Chao, C., and Thomaz, A. (2010). Designing interactions for robot active learners. *IEEE TAMD*, 2(2):108–118.

Cangelosi, A., Metta, G., Sagerer, G., Nolfi, S., Nehaniv, C., Fischer, K., Tani, J., Belpaeme, T., Sandini, G., Nori, F., et al. (2010). Integration of action and language knowledge: A roadmap for developmental robotics. *IEEE Transactions on Autonomous Mental Development*, 2(3):167–195.

Carey, S. (1978). The child as word learner. In Halle, M., Bresnan, J., and Miller, G. A., editors, *Linguistic theory and psychological reality*, pages 305–315. The MIT Press, Cambridge, MA.

Carey, S. (1991). Knowledge acquisition: Enrichment or conceptual change? In Carey, S. and Gelman, R., editors, *The Epigenesis of Mind: Essays on Biology and Cognition*, volume 41, pages 257–291. Erlbaum.

Carroll, J., Neitz, J., and Neitz, M. (2002). Estimates of l:m cone ratio from erg flicker photometry and genetics. *Journal of Vision*, 2(8).

Cassell, J. and Tartaro, A. (2007). Intersubjectivity in human-agent interaction. *Interaction Studies*, 8(3):391–410.

Chao, C., Cakmak, M., and Thomaz, A. (2010). Transparent active learning for robots. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2010)*, pages 317–324. IEEE.

Chao, C. and Thomaz, A. (2012). Timing in multimodal turn-taking interactions: Control and analysis using timed petri nets. *Journal of Human-Robot Interaction*, 1(1).

Chella, A., Frixione, M., and Gaglio, S. (1997). A cognitive architecture for artificial vision. *Artificial Intelligence*, 89(1):73 – 111.

Chella, A., Frixione, M., and Gaglio, S. (2000). Understanding dynamic scenes. *Artificial Intelligence*, 123(1):89 – 132.

Chiarello, C., Burgess, C., Richards, L., and Pollock, A. (1990). Semantic and associative priming in the cerebral hemispheres: Some words do, some words don't... sometimes, some places. *Brain and Language*, 38(1):75 – 104.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Praeger Publishers.

Cicerone, C. M. (1987). Constraints placed on color vision models by the relative numbers of different cone classes in human fovea centralis. *Die Farbe*, 34:59–66.

Clark, A. (2008). *Supersizing the mind: embodiment, action, and cognitive extension*. Oxford University Press, USA.

Clark, E. V. (1993). *The lexicon in acquisition*. Cambridge University Press, Cambridge.

Dautenhahn, K. (2007a). Methodology and themes of human-robot interaction: a growing research field. *International Journal of Advanced Robotic Systems*.

Dautenhahn, K. (2007b). Socially intelligent robots: dimensions of human–robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480):679–704.

Dautenhahn, K. and Billard, A. (1999). Studying robot social cognition within a developmental psychology framework. In *Third European Workshop on Advanced Mobile Robots (Eurobot '99)*, pages 187 –194.

Dautenhahn, K., Nehaniv, C. L., Walters, M. L., Robins, B., Kose-Bagci, H., Mirza, N. A., and Blow, M. (2009). KASPAR–a minimally expressive humanoid robot for human–robot interaction research. *Applied Bionics and Biomechanics*, 6(3-4):369–397.

Davidsson, P. (1996). *Autonomous Agents and the Concept of Concepts*. PhD thesis, Department of Computer Science, Lund University.

Delaunay, F. (2014). *A Rear-Projected Robotic Head for Social Human-Robot Interaction*. PhD thesis, Plymouth University, United Kingdom.

Delaunay, F., de Greeff, J., and Belpaeme, T. (2009). Towards retro-projected robot faces: an alternative to mechatronic and android faces. In *Proceedings of the International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Toyama, Japan.

Delaunay, F., de Greeff, J., and Belpaeme, T. (2010). A study of a retro-projected robotic face and its effectiveness for gaze reading by humans. In *Proceeding of the 5th ACM/IEEE international conference on Human-robot interaction*, pages 39–44, Osaka, Japan. ACM/IEEE.

de Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M. J., Voorspoels, W., and Storms, G. (2008). Exemplar by feature applicability matrices and other dutch normative data for semantic concepts. *Behavior research methods*, 40(4):1030–48.

Duda, R. O. and Hart, P. E. (1973). Pattern classification and scene analysis. *A Wiley-Interscience Publication, New York: Wiley, 1973*, 1.

Duffy, B. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, 42(3):177–190.

Dufour, R. and Kroll, J. (1995). Matching words to concepts in two languages: A test of the concept mediation model of bilingual representation. *Memory & Cognition*, 23:166–180.

van Eijck, J. and Zwarts, J. (2004). Formal concept analysis and prototypes. In Werner Kuhn, Martin Raubal, F. P. and Jonowicz, K., editors, *Workshop on the Potential of Cognitive Semantics for Ontologies*. University of Muenster.

Ekman, P. and Friesen, W. V. (1981). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Nonverbal communication, interaction, and gesture*, pages 57–106.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2):179 – 211.

Fairchild, M. (1998). *Color Appearance Models*. Addison-Wesley, Reading, MA.

Fairchild, M. D. (2005). *Color Appearance Models*. Wiley Interscience, Reading, MA, 2nd edition.

Fodor, J. (1998). *Concepts: Where cognitive science went wrong*. Oxford University Press, USA.

Fodor, J. A. (1981). *The present status of the innateness controversy*, pages 257–316. Number 1980. MIT Press.

Fong, T., Nourbakhsh, I., and Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3):143–166.

Frank, A. and Asuncion, A. (2010). UCI machine learning repository.

Franklin, A., Drivonikou, G. V., Bevis, L., Davies, I. R. L., Kay, P., and Regier, T. (2008). Categorical perception of color is lateralized to the right hemisphere in infants, but to the left hemisphere in adults. *Proceedings of the National Academy of Sciences*, 105(9):3221–3225.

French, R. (1999). Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning*, 29(2-3):131–163.

Ganter, B., Stumme, G., and Wille, R., editors (2005). *Formal Concept Analysis: Foundations and Applications*, volume 3626 of *Lecture Notes in Artificial Intelligence*, Heidelberg. Springer.

Gärdenfors, P. (2000a). Concept combination: a geometrical model. In Cavedon, L., Blackburn, P., Braisby, N., and Shimojima, A., editors, *Logic, Language and Computation*, volume 3, pages 129–146. CSLI Publications, Stanford, CA.

Gärdenfors, P. (2000b). *Conceptual Spaces: The Geometry of Thought*. the MIT Press, Cambridge, MA.

Gärdenfors, P. (2007). Representing actions and functional properties in conceptual spaces. In Ziemke, T., Zlatev, J., and Frank, R., editors, *Body, Language and Mind*, volume 1: Embodiment, pages 159–192. Mouton de Gruyter, Berlin.

Gauvain, M. (2001). *The social context of cognitive development*. The Guilford Press.

Gegenfurtner, K. R. and Sharpe, L. T., editors (1999). *Color Vision: From Genes to Perception*. Cambridge University Press, New York.

Gilbert, A. L., Regier, T., Kay, P., and Ivry, R. B. (2006). Whorf hypothesis is supported in the right visual field but not the left. *Proceedings of the National Academy of Sciences of the United States of America*, 103(2):489–494.

Glenberg, A. and Kaschak, M. (2002). Grounding language in action. *Psychonomic bulletin & review*, 9(3):558–565.

Goetz, J., Kiesler, S., and Powers, A. (2003). Matching robot appearance and behavior to tasks to improve human-robot cooperation. In *Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003. The 12th IEEE International Workshop on*, pages 55–60. IEEE.

Goldin-Meadow, S. (2003). *The resilience of language: What gesture creation in deaf children can tell us about how all children learn language*. Psychology Press, New York.

Goldin-Meadow, S., Gelman, S. A., and Mylander, C. (2005). Expressing generic concepts with and without a language model. *Cognition*, 96(2):109 – 126.

Goldstone, R. L. and Kersten, A. (2003). *Concepts and Categorization*. John Wiley & Sons, Inc.

Gonsior, B., Sosnowski, S., Mayer, C., Blume, J., Radig, B., Wollherr, D., and Kuhnlenz, K. (2011). Improving aspects of empathy and subjective performance for hri through mirroring facial expressions. In *RO-MAN, 2011 IEEE*, pages 350–356. IEEE.

Goodrich, M. A. and Schultz, A. C. (2007). Human-robot interaction: a survey. *Found. Trends Hum.-Comput. Interact.*, 1(3):203–275.

Gordon, P. (2004). Numerical cognition without words: Evidence from amazonia. *Science*, 306(5695):496–499.

de Greeff, J., Baxter, P., Wood, R., and Belpaeme, T. (2012a). From Penguins to Parakeets: a developmental approach to modelling conceptual prototypes. In Szufnarowska, J., editor, *Proceedings of the Post-Graduate Conference on Robotics and Development of Cognition*, pages 8–11, Lausanne, Switzerland.

de Greeff, J. and Belpaeme, T. (2011a). Coordination of meaning within different embodiments through linguistic interactions. In *Alife Approaches to Artificial Language Evolution (AAALE), workshop at the 20th European Conference on Artificial Life (ECAL)*, Paris, France.

de Greeff, J. and Belpaeme, T. (2011b). The development of shared meaning within different embodiments. In Triesch, J., editor, *Proceedings of the joint International Conference on Developmental Learning (ICDL) & Epigenetic Robotics 2011*, Frankfurt, Germany. IEEE.

de Greeff, J., Delaunay, F., and Belpaeme, T. (2009a). Concept acquisition through linguistic human-robot interaction. In *Proceedings of the International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Toyama, Japan.

de Greeff, J., Delaunay, F., and Belpaeme, T. (2009b). Human-robot interaction in concept acquisition: a computational model. In Triesch, J. and Zhang, Z., editors, *IEEE International Conference on Development and Learning (ICDL 2009)*, pages 1–6, Shanghai, China. IEEE.

de Greeff, J., Delaunay, F., and Belpaeme, T. (2012b). Active robot learning with human tutelage. In *Proceedings of the joint International Conference on Developmental Learning (ICDL) & Epigenetic Robotics 2012*, pages 1–6, San Diego, USA. IEEE.

Groves, R. M., Floyd, Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2004). *Survey Methodology*. Wiley Series in Survey Methodology. John Wiley and Sons, New York.

Hampton, J. (1995). Testing the prototype theory of concepts. *Journal of Memory and Language*, 34(5):686 – 708.

Hampton, J. A. and Gardiner, M. M. (1983). Measures of internal category structure: A correlational analysis of normative data. *British Journal of Psychology*, 74(4):491–516.

Han, J., Campbell, N., Jokinen, K., and Wilcock, G. (2012). Investigating the use of non-verbal cues in human-robot interaction with a Nao robot. In *Cognitive Infocommunications (CogInfoCom), 2012 IEEE 3rd International Conference on*, pages 679–683.

Hansen, D. W. and Ji, Q. (2010). In the eye of the beholder: A survey of models for eyes and gaze. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(3):478–500.

Harnad, S., editor (1987). *Categorical Perception: The Groundwork of Cognition*. Cambridge University Press.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. The MIT Press, Cambridge, Massachusetts.

Hauser, M. D., Chomsky, N., and Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598):1569–1579.

Hebb, D. (1949). *The organization of behavior: A neuropsychological theory*. Wiley, New York.

Hinde, R. A. (1972). *Non-verbal communication*. Cambridge University Press.

Hofer, H., Carroll, J., Neitz, J., Neitz, M., and Williams, D. R. (2005). Organization of the human trichromatic cone mosaic. *The Journal of Neuroscience*, 25(42):9669–9679.

Ishihara, H., Yoshikawa, Y., and Asada, M. (2011). Realistic child robot "affetto" for understanding the caregiver-child attachment relationship that guides the child development. In *Development and Learning (ICDL), 2011 IEEE International Conference on*, volume 2, pages 1–5. IEEE.

Ishihara, S. (2001). *Ishihara's tests for colour deficiency*. Kanehara and co., Tokyo. The series of plates designed as a test for colour deficiency, 24 plates edition.

James, W. (1890). *The Principles of Psychology*, volume 1. Henry Holt, New York.

John, O., Naumann, L., and Soto, C. (2008). Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues. In John, O., Robins, R., and Pervin, L., editors, *SAGE Handbook of Personality Theory and Assessment*, pages 114–158. Guilford Press.

de Jong, E. (1999). Autonomous concept formation. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence IJCAI'99*, pages 344–349. Morgan Kaufmann.

Jordan, M. (1997). Serial order: A parallel distributed processing approach. *Advances in psychology*, 121:471–495.

Kahn, P. H., Ishiguro, H., Friedman, B., Kanda, T., Freier, N. G., Severson, R. L., and Miller, J. (2007). What is a human? toward psychological benchmarks in the field of humanrobot interaction. *Interaction Studies*, 8(3):363–390.

Kanda, T., Hirano, T., Eaton, D., and Ishiguro, H. (2004). Interactive robots as social partners and peer tutors for children: A field trial. *Human-Computer Interaction*, 19(1):61–84.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., and Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):881–892.

Kay, P. and Kempton, W. (1984). What is the sapir-whorf hypothesis? *American Anthropologist*, 86(1):pp. 65–79.

Kidd, C. D. and Breazeal, C. (2005). Human-robot interaction experiments: Lessons learned. In *Proceeding of AISB*, volume 5, pages 141–142.

Kim, E. S., Paul, R., Shic, F., and Scassellati, B. (2012). Bridging the research gap: Making HRI useful to individuals with autism. *Journal of Human-Robot Interaction*, 1(1).

Knox, W., Glass, B., Love, B., Maddox, W., and Stone, P. (2012). How humans teach agents. *International Journal of Social Robotics*, pages 1–13.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69.

Kohonen, T. (1984). *Self-Organization and Associative Memory*. Springer Verlag, Berlin.

Kotsiantis, S., Zaharakis, I., and Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Frontiers in Artificial Intelligence and Applications*, 160:3.

Kraut, R. (2011). Plato. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Fall 2011 edition.

Krcmar, M., Grela, B., and Lin, K. (2007). Can toddlers learn vocabulary from television? An experimental approach. *Media Psychology*, 10(1):41–63.

Kuhn, T. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, Illinois.

Lakoff, G. (2008). The neural theory of metaphor. In Gibbs, R. J., editor, *The Cambridge Handbook of Metaphor and Thought*, pages 17–38. Cambridge University Press.

Landauer, T. and Dumais, S. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211.

Langton, S. R., Watt, R. J., and Bruce, V. (2000). Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Science*, 4(2):50–59.

Leite, I., Martinho, C., and Paiva, A. (2013). Social robots for long-term interaction: A survey. *International Journal of Social Robotics*, pages 1–18.

Ling, B. Y. and Dain, S. J. (2008). Color vision in children and the lanthony new color test. *Visual Neuroscience*, 25(03):441–444.

Lloyd, S. (1982). Least squares quantization in PCM. *Information Theory, IEEE Transactions on*, 28(2):129–137.

Lormand, E. (1990). Framing the frame problem. *Synthese*, 82:353–374.

Lund, K., Burgess, C., and Atchley, R. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th annual conference of the Cognitive Science Society*, volume 17, pages 660–665.

Lungarella, M., Metta, G., Pfeifer, R., and Sandini, G. (2003). Developmental robotics: a survey. *Connection Science*, 15(4):151–190.

Lupyan, G. (2006). Labels facilitate learning of novel categories. In *Proceedings of the 6th International Conference on the Evolution of Language*, pages 190–197.

Lupyan, G., Rakison, D. H., and McClelland, J. L. (2007). Language is not just for talking - redundant labels facilitate learning of novel categories. *Psychological Science*, 18(12):1077–1083.

Machery, E. (2009). *Doing without concepts*. Oxford University Press, USA.

MacWhinney, B. (2000). *The CHILDES project: tools for analyzing talk*. Lawrence Erlbaum, Mahwah.

Majid, A., Bowerman, M., Kita, S., Haun, D. B. M., and Levinson, S. C. (2004). Can language restructure cognition? The case for space. *Trends in Cognitive Sciences*, 8(3):108 – 114.

Mandler, J. (1997). Development of categorization: Perceptual and conceptual categories. In Bremner, G., Slater, A., and Butterworth, G., editors, *Infant development: Recent advances*, pages 163–189. Psychology Press, Hove, UK.

Margolis, E. and Laurence, S. (1999). *Concepts: Core Readings*. MIT Press.

Massé, A. B., Chicoisne, G., Gargouri, Y., Harnad, S., Picard, O., and Marcotte, O. (2008). How is meaning grounded in dictionary definitions? *CoRR*, abs/0806.3710.

McCarthy, J. and Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. In Meltzer, B. and Michie, D., editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press.

McClelland, J. L. and Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, 88(5):375–407.

Medin, D. L. and Schaffer, M. M. (1978). Context theory of classification learning. *Psychological review*, 85:207–238.

Mervis, C. and Crisafi, M. (1982). Order of acquisition of subordinate-, basic-, and superordinate-level categories. *Child Development*, pages 258–266.

Metta, G., Sandini, G., Vernon, D., Natale, L., and Nori, F. (2008). The iCub humanoid robot: an open platform for research in embodied cognition. In *Proceedings of the 8th workshop on performance metrics for intelligent systems*, pages 50–56. ACM.

Minato, T., Yoshikawa, Y., Noda, T., Ikemoto, S., Ishiguro, H., and Asada, M. (2007). Cb2: A child robot with biomimetic body for cognitive developmental robotics. In *Humanoid Robots, 2007 7th IEEE-RAS International Conference on*, pages 557–562. IEEE.

Mitchell, T. M. (1997). *Machine learning*. McGraw Hill series in computer science. McGraw-Hill.

Miyauchi, D., Nakamura, A., and Yoshinori, K. (2005). Bidirectional eye contact for human-robot communication. *IEICE transactions on information and systems*, 88(11):2509–2516.

Miyauchi, D., Sakurai, A., Nakamura, A., and Kuno, Y. (2004). Active eye contact for human-robot communication. In *CHI'04 extended abstracts on Human factors in computing systems*, pages 1099–1102. ACM.

Mori, M. (1970). The uncanny valley. *Energy*, 7(4):33–35.

Morimoto, C. and Mimica, M. (2005). Eye gaze tracking techniques for interactive applications. *Computer Vision and Image Understanding*, 98(1):4–24.

Morse, A., de Greeff, J., Belpaeme, T., and Cangelosi, A. (2010). Epigenetic robotics architecture (ERA). *IEEE Transactions on Autonomous Mental Development*, 2(4):325–339.

Munakata, Y. and Pfaffly, J. (2004). Hebbian learning and development. *Developmental Science*, 7(2):141–148.

Murphy, G. and Brownell, H. (1985). Category differentiation in object recognition: typicality constraints on the basic category advantage. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11(1):70.

Murphy, G. and Lassaline, M. (1997). Hierarchical structure in concepts and the basic level of categorization. *Knowledge, concepts, and categories*, pages 93–131.

Murphy, G. L. (2002). *The Big Book of Concepts*. MIT Press.

Murphy, G. L. and Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92(3):289–316.

Nagai, Y., Asada, M., and Hosoda, K. (2006). Learning for joint attention helped by functional development. *Advanced Robotics*, 20(10):1165–1181.

Neely, J. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, 106(3):226.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In Besner, D. and Humphreys, G. W., editors, *Basic processes in reading: Visual word recognition*, pages 265–335. Lawrence Erlbaum Associates, Hillsdale.

Neitz, J., Carroll, J., Yamauchi, Y., Neitz, M., and Williams, D. R. (2002). Color perception is mediated by a plastic neural mechanism that is adjustable in adults. *Neuron*, 35(4):783 – 792.

Nilsson, N. J. (1984). Shakey the robot. Technical Report 323, AI Center, SRI International, 333 Ravenswood Ave., Menlo Park, CA 94025.

Nishio, S., Ishiguro, H., and Hagita, N. (2007). Geminoid: Teleoperated android of an existing person. *Humanoid robots-new developments. I-Tech*, 14.

Nosofsky, R. and Zaki, S. (2002). Exemplar and prototype models revisited: Response strategies , selective attention , and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(5):924.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology-General*, 115(1):39–57.

O'Connor, C. M., Cree, G. S., and McRae, K. (2009). Conceptual Hierarchies in a Flat Attractor Network: Dynamics of Learning and Computations. *Cognitive Science*, 33(4):665–708.

Oh, J.-H., Hanson, D., Kim, W.-S., Han, Y., Kim, J.-Y., and Park, I.-W. (2006). Design of android type humanoid robot Albert HUBO. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, pages 1428–1433. IEEE.

O'Hanlon, C. and Roberson, D. (2007). What constrains children's learning of novel shape terms? *Journal of Experimental Child Psychology*, 97:138–148.

Oliphant, M. (1999). The learning barrier: Moving from innate to learned systems of communication. *Adaptive Behavior*, 7(3/4).

Onuki, T., Ishinoda, T., Kobayashi, Y., and Kuno, Y. (2013). Designing robot eyes for gaze communication. In *Frontiers of Computer Vision, (FCV), 2013 19th Korea-Japan Joint Workshop on*, pages 97–102.

O'Reilly, R. (1998). Six principles for biologically based computational models of cortical cognition. *Trends in cognitive sciences*, 2(11):455–462.

Osherson, D. N. and Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition*, 9(1):35–58.

Otake, S. and Cicerone, C. M. (2000). L and m cone relative numerosity and red–green opponency from fovea to midperiphery in the human retina. *J. Opt. Soc. Am. A*, 17(3):615–627.

Oudeyer, P.-Y. and Delaunay, F. (2008). Developmental exploration in the cultural evolution of lexical conventions. In *Proceedings of the 8th International Conference on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, Lund, Brighton. Lund University Cognitive Studies.

Page, M. (2000). Connectionist modelling in psychology: A localist manifesto. *Behavioral and Brain Sciences*, 23(04):443–467.

Pfeifer, R., Bongard, J., and Grand, S. (2007). *How the body shapes the way we think: a new view of intelligence*. The MIT Press.

Pfeifer, R. and Scheier, C. (2001). *Understanding intelligence*. The MIT Press.

Piaget, J. and Cook, M. (1953). *The origin of intelligence in the child*. Routledge & Kegan Paul Limited.

Plunkett, K., Hu, J., and Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. *Cognition*, 106(2):665–681.

Posner, M. and Keele, S. (1968). On the genesis of abstract ideas. *Journal of experimental psychology*, 77(3p1):353.

Priss, U. (2006). Formal concept analysis in information science. *Annual review of information science and technology*, 40:521.

Puglisi, A., Baronchelli, A., and Loreto, V. (2008). Cultural route to the emergence of linguistic categories. *Proceedings of the National Academy of Sciences*, 105(23):7936–7940.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

Quinn, P. C. (2003). Concepts are not just for objects: Categorization of spatial relation information by infants. In Rakison, D. H. and Oakes, L. M., editors, *Early category and concept development: Making sense of the blooming, buzzing confusion*, pages 50–76. Oxford University Press, New York, NY, USA.

Quinn, P. C. and Eimas, P. D. (2000). The emergence of category representations during infancy: Are separate perceptual and conceptual processes required? *Journal of Cognition and Development*, 1(1):55–61.

Rakison, D. and Oakes, L. (2003). *Early category and concept development: Making sense of the blooming, buzzing confusion*. Oxford University Press, New York, NY, USA.

Rakison, D. H. (2003). Parts, motion, and the development of the animate-inanimate distinction in infancy. In Rakison, D. H. and Oakes, L. M., editors, *Early category and concept development: Making sense of the blooming, buzzing confusion*, pages 159–192. Oxford University Press, New York, NY, USA.

Reed, S. (1972). Pattern recognition and categorization. *Cognitive psychology*, 3(3):382–407.

Regier, T. and Kay, P. (2009). Language, thought, and color: Whorf was half right. *Trends in Cognitive Sciences*, 13(10):439–446.

Reilly, D., Cooper, L., and Elbaum, C. (1982). A neural model for category learning. *Biological cybernetics*, 45(1):35–41.

Rickard, J., Aisbett, J., and Gibbon, G. (2007). Knowledge representation and reasoning in conceptual spaces. In *IEEE Symposium on Foundations of Computational Intelligence (FOCI 2007)*, pages 583 –590.

Riek, L. D. (2012). Wizard of Oz studies in HRI: a systematic review and new reporting guidelines. *Journal of Human-Robot Interaction*, 1(1).

Roberson, D., Pak, H., and Hanley, J. R. (2008). Categorical perception of colour in the left and right visual field is verbally mediated: Evidence from korean. *Cognition*, 107(2):752 – 762.

Roberts, K. (1988). Retrieval of a basic-level category in prelinguistic infants. *Developmental Psychology*, 24(1):21.

Rogoff, B. (1990). *Apprenticeship in thinking: Cognitive development in social context.* Oxford University Press.

Rojas, R. (1996). *Neural networks: a systematic introduction.* Springer-Verlag, Berlin, New-York.

Roorda, A. and Williams, D. R. (1999). The arrangement of the three cone classes in the living human eye. *Nature*, 397:520–522.

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

Rosch, E. (1978). Principles of categorization. In Rosch, E. and Lloyd, B. B., editors, *Cognition and Categorization*, pages 27–48. Lawrence Erlbaum Associates, Hillsdale (NJ), USA.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., and Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8(3):382 – 439.

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, 4(3):328 – 350.

Rumelhart, D. E. and McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: Part 2. the contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, 89(1):60–94.

Sandini, G., Metta, G., and Vernon, D. (2004). Robotcub: an open framework for research in embodied cognition. In *Proceedings of the 4th IEEE/RAS International Conference on Humanoid Robots*, volume 1, pages 13 – 32.

Sankeralli, M. J. and Mullen, K. T. (1996). Estimation of the l-, m-, and s-cone weights of the postreceptoral detection mechanisms. *J. Opt. Soc. Am. A*, 13(5):906–915.

Sato, E., Yamaguchi, T., and Harashima, F. (2007). Natural interface using pointing behavior for human-robot gestural interaction. *Industrial Electronics, IEEE Transactions on*, 54(2):1105–1112.

Schyns, P. (1991). A modular neural network model of concept acquisition. *Cognitive Science*, 15(4):461–508.

Searle, J. (1980). Minds, brains and programs. *Behavioral and Brain Sciences*, 3(3):417–457.

Settles, B. (2009). Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Sharkey, A. and Sharkey, N. (2012). Granny and the robots: ethical issues in robot care for the elderly. *Ethics and Information Technology*, 14(1):27–40.

Sharkey, N. and Sharkey, A. (2010). The crying shame of robot nannies: an ethical appraisal. *Interaction Studies*, 11(2):161–190.

Sharpe, L. T., Stockman, A., Jägle, H., and Nathans, J. (1999). Opsin genes, cone photopigments, color vision, and color blindness. In Gegenfurtner, K. R. and Sharpe, L. T., editors, *Color Vision: From genes to perception.* Cambridge University Press, New York.

Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science*, 237(4820):1317–1323.

Shields, C. (2012). Aristotle. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Summer 2012 edition.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61(1):39–91.

Smith, A. D. M. (2005). The inferential transmission of language. *Adaptive Behavior*, 13(4):311–324.

Smith, E. E. (1995). Concepts and categorization. In Smith, E. E. and Osherson, D. N., editors, *Invitation to Cognitive Science: Thinking*, volume 3. MIT Press, Cambridge, MA.

Smith, E. E. and Medin, D. L. (1981). *Categories and concepts*, volume 4. Harvard University Press.

Smith, E. E. and Osherson, D. N. (1984). Conceptual combination with prototype concepts. *Cognitive Science*, 8(4):337 – 361.

Smith, E. E., Osherson, D. N., Rips, L. J., and Keane, M. (1988). Combining prototypes: A selective modification model. *Cognitive Science*, 12(4):485–527.

Smith, K., Smith, A., and Blythe, R. (2011). Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, 35(3):480–498.

Smith, K., Smith, A. D., Blythe, R. A., and Vogt, P. (2006). Cross-situational learning: a mathematical approach. In *Symbol grounding and beyond*, pages 31–44. Springer.

Smith, L. and Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3):1558–1568.

Solomon, S. G. and Lennie, P. (2007). The machinery of colour vision. *Nature reviews. Neuroscience*, 8(4):276–286.

Sowa, J. F., editor (1991). *Principles of semantic networks: explorations in the representation of knowledge*. Morgan Kaufmann series in representation and reasoning. Morgan Kaufmann.

Sowa, J. F. (1992). Semantic networks. *Encyclopedia of Artificial Intelligence*, 23(3):291–9.

Spranger, M., Loetzsch, M., and Pauw, S. (2010). Open-ended grounded semantics. In *Proceedings of the 2010 conference on ECAI 2010: 19th European Conference on Artificial Intelligence*, pages 929–934, Amsterdam, The Netherlands. IOS Press.

Steels, L. (1996a). Perceptually grounded meaning creation. In Tokoro, M., editor, *Proceedings of the International Conference on Multiagent Systems (ICMAS-96)*, pages 338–344, Menlo Park, CA. AAAI Press.

Steels, L. (1996b). Self-organizing vocabularies. In Langton, C. and Shimohara, T., editors, *Proceedings of the Conference on Artificial Life V (Alife V) (Nara, Japan)*, Cambridge, MA. The MIT Press.

Steels, L. (1997). The synthetic modeling of language origins. *Evolution of communication*, 1(1):1–34.

Steels, L. (1998). Structural coupling of cognitive memories through adaptive language games. In Pfeifer, R., Blumberg, B., Meyer, J.-A., and Wilson, S., editors, *From Animals to Animats 5: Proceedings of the Fifth International Conference on Simulation of Adaptive Behavior*, pages 263–269, Cambridge, MA. The MIT Press.

Steels, L. (1999). *The Talking Heads Experiment. Volume 1. Words and Meanings.* Laboratorium, Antwerpen. Limited pre-edition.

Steels, L. (2003). The evolution of communication systems by adaptive agents. In Alonso, E., Kudenko, D., and Kazakov, D., editors, *Adaptive Agents and Multi-Agent Systems*, volume 2636 of *Lecture Notes in Computer Science*, pages 559–559. Springer Berlin Heidelberg.

Steels, L. (2006). Experiments on the emergence of human communication. *Trends in Cognitive Sciences*, 10(8):347–349.

Steels, L. (2012). *Experiments in Cultural Language Evolution*, volume 3. John Benjamins.

Steels, L. and Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, 28(4):469–89. Target Paper, discussion 489-529.

Steels, L. and Kaplan, F. (2002a). Aibo's first words: The social learning of language and meaning. *Evolution of communication*, 4(1):3–32.

Steels, L. and Kaplan, F. (2002b). Bootstrapping grounded word semantics. In Briscoe, T., editor, *Linguistic evolution through language acquisition: formal and computational models*, pages 53–74. Cambridge University Press, Cambridge, UK.

Steels, L. and Spranger, M. (2008). Can body language shape body image? In Bullock, S., Noble, J., Watson, R., and Bedau, M. A., editors, *Artificial Life XI: Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems*, pages 577–584, Cambridge, MA. MIT Press.

Steels, L. and Vogt, P. (1997). Grounding adaptive language games in robotic agents. In Husbands, P. and Harvey, I., editors, *Proceedings of the Fourth European Conference on Artificial Life (ECAL'97), Complex Adaptive Systems*, Cambridge, MA. The MIT Press.

Steinfeld, A., Fong, T., Kaber, D., Lewis, M., Scholtz, J., Schultz, A., and Goodrich, M. (2006). Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pages 33–40. ACM.

Stockman, A. and Sharpe, L. T. (2000). The spectral sensitivities of the middle- and long-wavelength-sensitive cones derived from measurements in observers of known genotype. *Vision research*, 40(13):1711–1737.

Stramandinoli, F., Marocco, D., and Cangelosi, A. (2012). The grounding of higher order concepts in action and language: A cognitive robotics model. *Neural Networks*, 32(0):165 – 173.

Sturges, J. and Whitfield, T. A. (1995). Locating basic colours in the Munsell space. *COLOR Research and Application*, 20(6):364–376.

Takáč, M. (2008). Autonomous construction of ecologically and socially relevant semantics. *Cognitive Systems Research*, 9(4):293 – 311.

Takayama, L., Groom, V., and Nass, C. (2009). I'm sorry, dave: I'm afraid I won't do that: social aspects of human-agent conflict. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2099–2108. ACM.

Tan, L. H., Chan, A. H. D., Kay, P., Khong, P.-L., Yip, L. K. C., and Luke, K.-K. (2008). Language affects patterns of brain activation associated with perceptual decision. *Proceedings of the National Academy of Sciences*, 105(10):4004–4009.

Tanaka, F. and Matsuzoe, S. (2012). Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning. *Journal of Human-Robot Interaction*, 1(1).

Tanaka, J. and Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive psychology*, 23(3):457–482.

Tanasescu, V. (2007). Spatial semantics in difference spaces. In Winter, S., Duckham, M., Kulik, L., and Kuipers, B., editors, *Spatial Information Theory*, volume 4736 of *Lecture Notes in Computer Science*, pages 96–115. Springer Berlin Heidelberg.

Thomaz, A. L. (2006). *Socially Guided Machine Learning*. PhD thesis, MIT.

Thomaz, A. L. and Breazeal, C. (2008). Experiments in socially guided exploration: lessons learned in building robots that learn with and without human teachers. *Connection Science*, 20(2-3):91–110.

Tomasello, M. (1992). The social bases of language acquisition. *Social Development*, 1(1):67–87.

Tomasello, M. (1995). Joint attention as social cognition. In Moore, C. and Dunham, P., editors, *Joint attention Its origins and role in development*, pages 103–130. Lawrence Erlbaum, Hillsdale, NJ.

Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Harvard University Press, Cambridge, MA.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press, Cambridge, MA.

Towell, G. G. and Shavlik, J. W. (1994). Knowledge-based artificial neural networks. *Artificial Intelligence*, 70(1):119 – 165.

Trevarthen, C. (1998). The concept and foundations of infant intersubjectivity. In Braten, S., editor, *Intersubjective Communication and Emotion in Early Ontogeny*, pages 15–46. Cambridge University Press, Cambridge, UK.

Turing, A. (1950). Computing machinery and intelligence. *Mind*, 59(236):433–460.

Uzgalis, W. (2010). John locke. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Winter 2010 edition.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA.

Vogt, P. (2004). Minimum cost and the emergence of the Zipf-Mandelbrot law. In *Artificial Life IX: Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems*.

Vogt, P. (2005). The emergence of compositional structures in perceptually grounded language games. *Artificial Intelligence*, 167(1-2):206–242.

Vogt, P. and Divina, F. (2007). Social symbol grounding and language evolution. *Interaction Studies*, 8(1):31–52.

Vogt, P. and Haasdijk, E. (2010). Modeling social learning of language and skills. *Artificial Life*, 16(4):289–309.

Voorspoels, W., Storms, G., and Vanpaemel, W. (2011). Representation at different levels in a conceptual hierarchy. *Acta psychologica*, 138(1):11–18.

Vygotsky, L. (1964). Thought and language. *Annals of Dyslexia*, 14(1):97–98.

Wada, K., Shibata, T., Musha, T., and Kimura, S. (2005). Effects of robot therapy for demented patients evaluated by EEG. In *Intelligent Robots and Systems, 2005. (IROS 2005). 2005 IEEE/RSJ International Conference on*, pages 1552–1557.

Walters, M. L., Woods, S., Koay, K. L., and Dautenhahn, K. (2005). Practical and methodological challenges in designing and conducting human-robot interaction studies. pages 110–120.

Webster, M. A., Miyahara, E., Malkoc, G., and Raker, V. E. (2000). Variations in normal color vision. i. cone-opponent axes. *J. Opt. Soc. Am. A*, 17(9):1535–1544.

Weiss, A., Bernhaupt, R., Lankes, M., and Tscheligi, M. (2009). The usus evaluation framework for human-robot interaction. In *AISB2009: proceedings of the symposium on new frontiers in human-robot interaction*, volume 4, pages 11–26. Citeseer.

Weiss, A., Igelsböck, J., Tscheligi, M., Bauer, A., Kühnlenz, K., Wollherr, D., and Buss, M. (2010). Robots asking for directions: the willingness of passers-by to support robots. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*, pages 23–30, Piscataway, NJ, USA. IEEE Press.

Wellens, P., Loetzsch, M., and Steels, L. (2008). Flexible word meaning in embodied agents. *Connection Science*, 20(2):173–191.

Wennekers, T. (2009). On the natural hierarchical composition of cliques in cell assemblies. *Cognitive Computation*, 1(2):128–138.

Whorf, B. and Carroll, J. (1956). *Language, thought, and reality: Selected writings*, volume 5. the MIT Press.

Wiemer-Hastings, K., Krug, J., and Xu, X. (2001). Imagery, context availability, contextual constraint, and abstractness. In *Proceedings of the 23rd annual conference of the cognitive science society*, pages 1134–1139.

Wiener, N. (1948). *Cybernetics; or, Control and communication in the animal and the machine*. New York: John Wiley.

Wille, R. (1982). Restructuring lattice theory: an approach based on hierarchies of concepts. In Rival, I., editor, *Ordered Sets*, pages 445–470. Proc. Nato Advanced Study Institute, vol. 1. Reidel, Dordrecht.

Williams, D. R., MacLeod, D. I., and Hayhoe, M. M. (1981a). Foveal tritanopia. *Vision Research*, 21(9):1341 – 1356.

Williams, D. R., MacLeod, D. I., and Hayhoe, M. M. (1981b). Punctate sensitivity of the blue-sensitive mechanism. *Vision Research*, 21(9):1357 – 1375.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, 9:625–636.

Wittgenstein, L. (1953). *Philosophical Investigations*. Basil Blackwell, Oxford.

Woods, S., Walters, M., Koay, K., and Dautenhahn, K. (2006). Comparing human robot interaction scenarios using live and video based methods: towards a novel methodological approach. In *Advanced Motion Control, 2006. 9th IEEE International Workshop on*, pages 750–755.

Wyszecki, G. and Stiles, W. (1982). *Color Science: Concepts and Methods, Quantitative Data and Formulae*. John Wiley and sons, New York, 2nd edition. Reprinted in 2000.

Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition*, 85(3):223 – 250.

Yang, H.-D., Park, A.-Y., and Lee, S.-W. (2007). Gesture spotting and recognition for human-robot interaction. *Robotics, IEEE Transactions on*, 23(2):256–270.

Yoo, D. H. and Chung, M. J. (2005). A novel non-intrusive eye gaze estimation using cross-ratio under large head motion. *Comput. Vis. Image Underst.*, 98(1):25–51.

Yoshikawa, Y., Shinozawa, K., Ishiguro, H., Hagita, N., and Miyamoto, T. (2006). Responsive robot gaze to interaction partner. In *Robotics: Science and Systems*.