Referring Expression Generation in Situated Interaction

Dissertation zur Erlangung des Grades

Doktor der Ingenieurwissenschaften (Dr.-Ing.)

des Fachbereichs 3 Mathematik/Informatik

der Universität Bremen

Vivien Mast

November 2016

Gutachter 1: Prof. John A. Bateman, PhD

Gutachterin 2: Prof. Dr. Tanja Schultz

Gutachter 3: Prof. Kees van Deemter, PhD

Zusammenfassung

Wenn wir Maschinen schaffen wollen, die Menschen erfolgreich bei Alltagsverrichtungen unterstützen, ist natürliche und reibungslose Kommunikation essentiell. Dies schließt insbesondere die Kommunikation über die räumliche Umgebung und die darin enthaltenen Objekte mit ein. Bei der Objektreferenz in situierter Interaktion führen jedoch Perspektivunterschiede und die unterschiedliche Wahrnehmung von Menschen und ihren künstlichen Kommunikationspartnern zu perzeptuellen und konzeptuellen Diskrepanzen, die die Kommunikation behindern können.

Während Menschen in der Kommunikation Konzeptualisierungen flexibel anpassen können, um Verständigung mit einem Kommunikationspartner zu erlangen und somit Unterschiede zwischen Weltsichten zu überbrücken, verwenden die meisten gegenwärtigen Ansätze zur Objektreferenz binäre Wissensrepräsentation mit Wahrheitswerten. Diese Ansätze basieren auf der Annahme, dass a priori feststellbar ist, ob eine bestimmte Eigenschaft auf ein gegebenes Objekt zutrifft oder nicht. Dabei verzichten Sie auf die Möglichkeit, durch eine flexiblere Herangehensweise Anpassung und Verständigung im Dialog zu erreichen, und damit Missverständnisse zu vermeiden und eine reibungslosere und natürlichere Kommunikation zu ermöglichen.

In dieser Dissertation argumentiere ich für einen Fokus auf das kollaborative Wesen von Referenz und die zielorientierte, flexible Natur von Konzeptualisierung. Auf dieser Grundlage präsentiere ich PRAGR, den probabilistischen Mechanismus für Objektbenennung und Verständigung (**P**robabilistic **R**eference **A**nd **GR**rounding mechanism), der dazu beiträgt, die Lücke zwischen menschlichen und künstlichen Kommunikationsteilnehmern in situierter Interaktion zu überbrücken. PRAGR geht über die meist für die Objektreferenz verwendete klassische, wahrheitswertbasierte Wissensrepräsentation hinaus, und verwendet stattdessen flexible Konzeptualisierung auf der Grundlage vager Eigenschaftsmodelle und situativem Kontext mit dem Ziel der Maximierung der Wahrscheinlichkeit kommunikativen Erfolgs. Zu diesem Zweck verwendet PRAGR die Kernkonzepte der Akzeptabilität (acceptability), Unterscheidungskraft (discriminatory power), und Ange-

messenheit (appropriateness), die ein Mittel zur Bewertung referierender Ausdrücke für die Generierung und Interpretation von Objektreferenz liefern, und die inhärent auf konzeptueller Vagheit basieren.

Ich demonstriere mit dieser Arbeit, dass PRAGR in der Lage ist, zahlreiche unterschiedlich strukturierte konzeptuelle Domänen zu integrieren, darunter Gradadjektive, Farbe, Form, Richtungsrelationen, und Richtungsregionen. Desweiteren zeige ich, dass PRAGR geeignet ist, die wichtigsten Herausforderungen der Generierung referierender Ausdrücke in integrierter Weise zu meistern, insbesondere Gradadjektive, referierende Ausdrücke mit räumlichen Relationen, und Salienzeffekte.

Ein Schwerpunkt ist hierbei die Integration räumlicher Relationen in die Generierung referierender Ausdrücke, zu der diese Dissertation einen besonderen Beitrag leistet. Im Gegensatz zum Großteil bestehender Arbeiten betrachte ich in dieser Dissertation Objektreferenz mit räumlichen Relationen als Unterstützung visueller Suche und integriere Erkenntnisse der Forschung zur Wahl von Referenzobjekten. Somit liefert die vorliegende Dissertation den bislang einzigen Mechanismus zur Generierung referierender Ausdrücke mit einer hoch differenzierten Auswahl von Referenzobjekten, die die Einflussfaktoren der Lokalisierbarkeit des Referenzobjekts, der Suchraumoptimierung, und der kommunikativen Kosten integriert. In diesem Kontext stelle ich auch die Erweiterung von PRAGR zur Berücksichtigung von Salienz vor, und einen Suchalgorithmus für PRAGR, der die Komplexitätsprobleme bewältigt, die durch die Kombination von Vagheit und räumlichen Relationen in der Objektreferenz entstehen.

Um die Nützlichkeit von PRAGR für situierte Objektreferenz und seine Fähigkeit zum Umgang mit einer Vielzahl unterschiedlicher Eigenschaftsmodelle zu demonstrieren, präsentiere ich drei empirische Evaluationsstudien, die sowohl Mensch-Roboter Interaktion als auch Roboter-Roboter Interaktion beinhalten. Die Experimente zeigen, dass PRAGR in der Lage ist, von Menschen produzierte referierende Ausdrücke unter Bedingungen der perzeptuellen Abweichung mit hoher Treffgenauigkeit zu interpretieren, und selbst referierende Ausdrücke zu generieren, die von Menschen gut verstanden werden. Des Weite-

ren zeigen die Evaluationsstudien, dass die Verwendung vager Eigenschaftsmodelle, insbesondere auf Seiten des Hörers, die Erfolgsquote in Roboter-Roboter und Mensch-Roboter Interaktion verbessert.

Schließlich zeige ich in dieser Dissertation Wege auf, wie PRAGR referenzielle Verständigungsdialoge unterstützen kann und präsentiere die Integration von PRAGR in das DAISIE Framework für Dialogsysteme am Beispiel eines einfachen Szenarios für referenzielle Verständigungsdialoge.

Abstract

If we want to enable artificial agents to successfully support humans in everyday life, natural and smooth communication is the key. In particular, communication about the spatial environment and the objects contained therein is crucial. However, when referring to objects in situated interaction, the difference in perspective and perceptual ability between humans and artificial agents leads to perceptual and conceptual mismatch which may hinder communication.

While humans are capable of flexibly adapting conceptualisations in communication, thus bridging gaps between individual views of the world to reach a mutual understanding, most current frameworks for reference handling are based on binary truth-theoretic knowledge representation. These approaches rest on the assumption that it is possible to unambiguously determine a priori whether a certain property is true of a given object or not, thus forgoing the opportunity to use a more flexible approach which may aid adaptation and grounding in dialogue and enable more smooth and natural communication while avoiding misunderstandings.

In this thesis, I argue for a perspective on reference which emphasises the collaborative nature of reference, and the goal-oriented, flexible nature of conceptualisation. Based on this foundation, I present the **P**robabilistic **R**eference **A**nd **GR**ounding mechanism (PRAGR) which can help bridge the gap between human and artificial communicators in situated interaction. PRAGR goes beyond the classic, truth-theoretic knowledge representation typically used for Referring Expression Generation (REG) and instead uses flexible concept assignment based on vague property models and situational context in order to maximise the chance of communicative success. To this purpose, PRAGR uses the core concepts of acceptability, discriminatory power, and appropriateness which provide a means for evaluating referring expressions for the purpose of REG and reference resolution and encompass conceptual vagueness as an inherent characteristic.

I demonstrate that PRAGR is capable of dealing with several property domains with different internal structures, covering the following domains: graded adjectives, colour, shape, projective terms, and pro-

jective regions. Further, I show that PRAGR is fit to handle in an integrated fashion the most relevant REG challenges, in particular graded properties, spatial relations, and salience effects in REG.

A focus of this thesis is the integration of spatial relations into REG where I propose a unique position in treating REG with relations as aiding visual search, and integrate findings on reference object selection, thus providing the first REG system which is capable of highly sophisticated reference object selection integrating the influence factors of reference object locatability, search space optimisation, and communication cost. In this context, I also present an extension of PRAGR to handle salience, and a search algorithm for PRAGR which overcomes the complexity issues raised by combining vagueness and spatial relations in REG.

In order to demonstrate the usefulness of PRAGR for situated referential communication and its ability to handle the interaction of a variety of property models, I present three empirical evaluation studies, covering both robot-robot and human-robot communication. The experiments show that PRAGR is capable of understanding human-produced referring expressions with a high degree of accuracy under conditions of perceptual deviation, and can generate referring expressions which are easily understood by human subjects. Further, I show that using vague property models, in particular on the side of the listener, improves task success in robot-robot and human-robot communication under conditions of perceptual deviation.

Finally, I discuss the ways in which PRAGR can support referential grounding dialogues, and present the integration of PRAGR with the DAISIE dialogue system framework and architecture for a simple referential grounding dialogue scenario.

Acknowledgements

"You see, Momo (...) it's like this. Sometimes, when you've a very long street ahead of you, you think how terribly long it is and feel sure you'll never get it swept (...) And then you start to hurry. (...) You work faster and faster, and everytime you look up there seems to be just as much left to sweep as before, and you try even harder, (...) and you panic, and in the end you're out of breath and have to stop – and still the street stretches away in front of you. That's not the way to do it. (...) You must never think of the whole street at once, understand? You must only concentrate on the next step, the next breath, the next stroke of the broom, and the next, and the next. Nothing else. That way you enjoy your work, which is important, because then you make a good job of it. And that's how it ought to be. (...) And all at once, before you know it, you find you've swept the whole street clean, bit by bit. What's more, you aren't out of breath. (...) That's important, too[.]" (Michael Ende: Momo)

The major challenge of the PhD enterprise is not the intellectual challenge, though that is also huge. The major challenge is facing that never-ending street that is a PhD thesis, and then making the first step towards cleaning it. And then another. And another, again and again, every day. I want to thank those who helped me clean the street that is my PhD thesis, but more than that I want to thank those who were there in the moments when I was out of breath, staring at the seemingly endless street stretching in front of me. Who helped me focus on the very next step, the very next sweep of the broom. Who helped me get to the end of this endeavour without getting out of breath.

Thanks go to my mentors and supervisors who have provided support

and guidance on the way. Thank you to my thesis supervisor John Bateman for supporting me in the development of my ideas even if they didn't always coincide with your research focus, and for having my back on the important things. Thank you also to Kees van Deemter for your inspiring work on the topic of reference, and for providing valuable feedback on publications leading to this thesis, and on an early version of the thesis. Thank you, Thora Tenbrink for your valuable feedback, and helping me keep focus when I was getting lost in details and perfectionism. A special Thank you to Diedrich Wolter for listening to my ideas in a time when I was stuck, and for helping me get them off the ground, and also for the productive collaboration over the years and many inspiring discussions and joint projects about getting together reference, communication, and spatial reasoning.

Thanks also to the Team of I5 and my friends and colleagues at the University of Bremen for the great time in Bremen, and exciting research with Rolland the wheelchair and the labyrinth of GW2. Special thanks to Daniel Couto Vale, for the inspiration, intense discussions, and productive collaboration, and also for your patience with my impatience, and excellent teamwork in the face of looming deadlines. Thank you, Zoe Falomir, for bringing your positive attitude into our joint work and giving me insights on the world of image processing. And thank you, Emma Bergmann and Benjamin Saade for being such excellent study buddies, for getting into R and the statistics world together with me, for your valuable feedback and intense moral support.

Thank you to the students who contributed to my thesis with their work. Thanks, Wang Tao for programming the GUI of the Dog Scene Demo. Thank you, Sandra Höfer, for providing an early version of what would later become my experimental GUI. And a special thanks to Elisa Vales and Rumiya Izgalieva who went above and beyond to support me with the organisation of experiments, taking photographs of cups, and various other tasks.

Thank you to my friends and family and all those other people who were there and made it all worthwhile. Thank you Uwe Staroske whose invaluable Yoga sessions helped me clear my head and keep a relaxed and calm attitude. Thanks to my sister Esther Münter for being calm and matter-of-fact in the

face of my crises. Thanks go to my mother, Patricia Mast for having raised me to trust in myself and to confront the challenges of life. Thank you for caring, thank you for worrying, thank you for being there and supportive no matter what. Thank you Anna Shadrova for your valuable feedback on early chapters of the thesis, for commiserations on the life of academia, and for just 'getting it' in so many different ways. Thank you Anna Steffen for being the excellent friend that you are, for your understanding and validation and support.

Thanks to the members of the Awkward Army (bear division, gemstone division, and all those other wonderful beings) who were generous with advice and jedi hugs and cheering me on, and helped me focus on the next sweep of the broom. Thanks for all the wisdom and compassion, the cheerfulness and positivity, the openness and awkwardness and vulnerability.

Clemens Haug, I don't even know how to begin thanking you for your incredible patience, support, and care in the form of a shoulder to lean on, delicious meals to keep me strong and happy, inspiring conversations and seemingly endless supply of love for the imperfect person that I am. Without you, I wouldn't be the person I am today, and this would all be meaningless.

Thanks to my cat Momo, who only ever sees the very next step, i.e., her next meal, and would never bother cleaning a street in the first place.

Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe. Alle Stellen, die ich wörtlich oder sinngemäß aus anderen Werken entnommen habe, habe ich unter Angabe der Quellen als solche kenntlich gemacht.

Vivien Mast
Leipzig, 28. November 2016

Related Publications

Parts of the results presented in this thesis have already been published in the form of book chapters, journal articles, and conference proceedings. The following publications are directly related to the thesis presented here.

Falomir, Z., Mast, V., Couto Vale, D., Museros, L., and Gonzalez-Abril, L. (2014). Towards a fuzzy colour descriptor sensitive to the context. In *XVI Jornadas de ARCA – Sistemas Cualitativos y sus Aplicaciones en Diagnosis, Robotica e Inteligencia Ambiental*.

Mast, V. (2016). Der Mythos der eindeutigen Beschreibung – Mehrdeutigkeit in der Objektbenennung als graduelles Phänomen. In Potysch, N. and Bauer, M., editors, *Deutungsspielräume. Mehrdeutigkeit als kulturelles Phänomen*. Peter Lang, Frankfurt a.M.

Mast, V., Couto Vale, D., and Falomir, Z. (2014a). Enabling grounding dialogues through probabilistic reference handling. In *RefNet Workshop on Psychological and Computational Models of Reference Comprehension and Production*, Edinburgh.

Mast, V., Couto Vale, D., Falomir, Z., and Elahi, M. F. (2014b). Referential grounding for situated human-robot communication. In Rieser, V. and Muller, P., editors, *Proceedings of SemDial 2014 - DialWatt*, pages 223–225.

Mast, V., Falomir, Z., and Wolter, D. (2016). Probabilistic reference and grounding with PRAGR for dialogues with robots. *Journal of Experimental and Theoretical Artificial Intelligence*.

Mast, V., Jian, C., and Zhekova, D. (2012). Elaborate descriptive information in indoor route instructions. In Miyake, N., Peebles, D., and Cooper, R., editors, *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, Austin, TX. Cognitive Science Society.

Mast, V. and Wolter, D. (2013a). Context and vagueness in REG. In *Proceedings of PRE-CogSci 2013*, Berlin.

Mast, V. and Wolter, D. (2013b). A probabilistic framework for object descriptions in indoor route instructions. In Tenbrink, T., Stell, J., Galton, A., and Wood, Z., editors, *Spatial Information Theory*, volume 8116 of *LNCS*, pages 185–204. Springer, Berlin/Heidelberg.

Mast, V., Wolter, D., Klippel, A., Wallgrün, J. O., and Tenbrink, T. (2014c). Boundaries and prototypes in categorizing direction. In *Spatial Cognition IX*, pages 92–107. Springer, Berlin/Heidelberg.

Contents

List of Figures

List of Tables

Glossary

Acceptability In this thesis, *Acceptability* has a two-fold meaning: On the one hand, the subjective (graded) willingness of the agent themselves to accept a given description as a valid conceptualisation of the object in question, and on the other hand, the intersubjective probability $P(D|x)$ that the listener would accept description D as true of object x.

Appropriateness The degree to which a description is deemed appropriate for describing a given object in a given context. Appropriateness is the weighted average of Discriminatory Power and Acceptability.

classic REG The approach to Referring Expression Generation which, in the tradition of Dale and Reiter (1995), is based on binary truth-conditional knowledge representation. classic REG has as its goal the generation of a distinguishing description for a target by selecting from a set of properties which are true of the target a subset of properties which are not all true for any of the distractors.

context set The set of objects that are perceptually or conceptually available in a scene and and can thus serve as potential targets of a referring expression. In contrast to a *distractor*, which refers to the set of objects that may be confused with the target, given a specific (preliminary) description, I use *context set* only to refer to the perceptual and conceptual situation and therefore the entirety of objects in the scene, independently of any descriptions.

Discriminatory Power In REG in general, *discriminatory power* is seen as the power of a description to discriminate the target object from

the distractor objects. In classic REG, this depends on the number of distractors which share all properties of the description. In this thesis, I use this term in a probabilistic sense, where *discriminatory power* refers to $P(x|D)$, the probability that the listener will identify the correct object as the target object of a given description.

distinguishing description A core concept of classic REG, a distinguishing description is a description which consists of a set of properties, all of which are true for the target, and for which there exists no distractor object for which all of the properties are true.

distractor An object which may be confused with the intended referent, given a (preliminary) description, due to being part of the same scene and sharing the properties used in the description. In contrast to the term *context set*, I use the term *distractor* to refer to the communicative situation. The term distractor therefore always presumes some (preliminary) description as the context in which the confusion may occur, even if this is the empty description.

intermediary description an *intermediary description* is any (partial or complete) description created during the REG process. Intermediary descriptions may not yet be fully evaluated and/or *resolved*, i.e., they may contain reference objects for which no description has been produced yet.

property model A *property model* is a function which assigns one or more properties an acceptability value $a \in [0, 1]$ for any given object, based on a given set of perceptual features of the object.

reference object The object which serves as the point of reference for a relational description of a given target object.

reference object selection The process or task of selecting a suitable reference object for a relational description of a given target object.

reference resolution The process or task of determining the intended referent of a given description. Not to be confused with resolution, a term

used in this thesis to denote the step of adding descriptions for objects contained in preliminary descriptions during REG.

referent The object which is being referred to with a given expression. In this thesis, the term *referent* always refers to the communicative situation and implies the presence of an RE. In contrast, I use the expression *target object* for pre-communicative settings when a speaker intends to refer to an object, but no specific RE is being considered.

relational referring expression Referring expression which contains at least one (spatial or other) relation.

resolution The process in which an unresolved description (UD) is expanded such that one unresolved reference object is described by adding a non-empty set of properties which describe this object.

resolved description A *resolved description* is an intermediary or final description for which all reference objects have been described. Thus, no further steps of resolution are necessary. This contrasts with an – always intermediary – unresolved description which contains reference objects which still lack a description.

salience The phenomenon that certain objects in a scene stand out perceptually or cognitively and are therefore more easily accessible than others, and communicatively preferred. In the context of this thesis, salience is used broadly to refer to some (graded) measure of how much a given object stands out in a given physical or linguistic context, as well as a measure of how much a given property stands out in a given context.

situated human-machine interaction The interaction between a human and a machine within a given physical setting. Typically, the machine in such an interaction is assumed to have some kind of sensors and actuators which enable it to perceive the situation and act within it. A prototypical example of situated human-machine interaction is human-robot interaction.

situated interaction An interaction which takes place within a given, usually physical, setting which plays a relevant role in the interaction. Situated interaction is characterised by the necessity to relate utterances not only to an abstract meaning, but to the specific properties and objects of the environment, and actual or hypothetical actions therein..

target object The object a speaker intends to refer to. In contrast to a referent, which is always the referent *of some expression* and therefore refers strictly to the communicative situation, in this thesis I use the term target object in a broader sense, including pre-communicative situations where no specific RE is present.

unresolved description An intermediary description in the REG process which has at least one reference object which has not yet been described.

Acronyms

AVS Attention Vector Sum Model.

FB Full Brevity Algorithm.

GDA Global Decision Approach.

GH Greedy Heuristic Algorithm.

HSL Hue, Saturation and Lightness.

IA Incremental Algorithm.

IDA Independent Decision Approach.

MDA Modular Decision Approach.

NLG Natural Language Generation.

NLU Natural Language Understanding.

NP Noun Phrase.

PRAGR **P**robabilistic **R**eference **A**nd **GR**ounding mechanism.

RD resolved description.

RE Referring Expression.

REG Referring Expression Generation.

RR reference resolution.

UD unresolved description.

Chapter 1

Introduction

Mary is resting on the sofa, relaxing after a long day at work. She decides to read the new book she got for her birthday last week, and asks her personal assistant robot Amanda: *Could you pass me that yellow book on my desk?* Amanda scans the desk and identifies that there are two books which seem to match that description. She asks: *Do you mean the one in front of the coffee cup?* Slightly annoyed, Mary replies: *No, not the green one, the yellow one.* Amanda understands and confirms: *Oh, okay. I'll get it.*, moves to the desk, grabs the book, and brings it to Mary.

interactants are jointly working towards a goal, or one interactant tries to instruct the other to achieve something for them.

In the example above, Mary is trying to instruct Amanda to pick up and bring her a specific book which she wants to read. Given that there are several books in the immediate vicinity, only one of which is relevant to Mary's goal, Mary engages in an act of referring.

She produces a Referring Expression (RE), with the goal of enabling Amanda to identify the object she has in mind (the book she wants to read), in order for her to perform some action on it (pick it up and bring it to her). Such acts of referring play a central role in human communication, and have been studied extensively (Krahmer and van Deemter, 2012, p. 174). Therefore, almost all implemented Natural Language Generation (NLG) systems have some kind of Referring Expression Generation (REG) component (Krahmer and van Deemter, 2012; Mellish et al., 2006).

Following Mary's statement, Amanda then generates an RE of her own (*Do you mean the one in front of the coffee cup?*). She in turn intends to enable Mary to identify the object she has in mind (the one she thinks Mary wants her to bring) in order to verify that they are talking about the same object. This process of grounding is a fundamental aspect of reference in human-human communication (Clark and Wilkes-Gibbs, 1986).

In referential grounding dialogues, rather than relying on fixed categorisations of objects, humans are capable of flexibly adapting their conceptualisation of objects in order to establish common ground with a communication partner, thus bridging gaps between individual views of the world to reach a mutual understanding (Clark and Wilkes-Gibbs, 1986; Garrod and Anderson, 1987; Pickering and Garrod, 2004). For example, the same object may be called *red* in one situation, and *orange* in another situation. This difference may be influenced by the communication partner, or by the presence of potential distractors – objects which may be confused with the intended target object due to sharing properties with it.

1.1 Goal of this Thesis

Enabling computational systems to bridge this gap is a big challenge: The gap to bridge is much larger than between average humans speaking the same language. The nature of the information an artificial system can gather, and the way this information is usually represented, is vastly different from the kind of information humans use for verbal interaction: Machines are good at handling exact quantitative information. With respect to spatial interaction, that might be angles, metric distances, or certain values of Hue, Saturation and Lightness (HSL) as a representation of colour. Humans, on the other hand, have vague, qualitative information about their surroundings – for example, we have colour terms which cannot be definitely ascribed to a certain range of hue, brightness and saturation values, as they may vary from person to person, and depending on lighting conditions and context. Humans use projective terms like LEFT or IN FRONT OF – terms which denote regions in space that have no clear-cut boundaries and cannot be defined by exact numbers and angles, as these may differ depending on the situational context.

Given the larger gap to bridge, there is need for reference handling systems which are to some degree flexible with respect to categorisation. Most current frameworks for reference handling are based on a binary truth-theoretic knowledge representation which assumes that for each object it can be unambiguously determined a priori whether a certain property is true of that object or not (Krahmer and van Deemter, 2012). In this thesis, I will argue that reference handling systems need to be integrated with a kind of knowledge representation which enables more flexible conceptualisations in adaptation to dialogue processes in order to enable computational systems to bridge the conceptual gap in reference handling.

Further, in order to be suited for application in situated human-machine interaction, a reference handling mechanism needs to be able to deal with a number of challenges which have been identified by the scientific community: (1) graded adjectives, which are used frequently in referential communication, pose problems for the crisp property models of classical approaches to REG, (2) the integration of spatial relations into REG algorithms and the

selection of suitable reference objects raises a number of issues which need to be addressed, (3) salience effects have been shown to influence the production and resolution of REs and therefore need to be taken into account, and (4) any proposed mechanism for REG needs to come with a search algorithm which allows finding an appropriate RE in a reasonable amount of time. While many of these challenges have been tackled by different authors, integrating the consideration of all these challenges into a single coherent mechanism for reference handling still poses a major difficulty (Krahmer and van Deemter, 2012).

In this thesis, I will present the **P**robabilistic **R**eference **A**nd **GR**ounding mechanism (PRAGR) mechanism which can help bridge the gap between human and artificial communicators in spatial interaction by going beyond the classic, truth-theoretic knowledge representation typically used for REG. Instead, PRAGR uses flexible concept assignment based on vague property models and situational context in order to maximise the chance of communicative success. I will demonstrate that PRAGR is capable of dealing with several property domains with different internal structures, and that it is fit to handle in an integrated fashion the most relevant REG challenges, in particular graded properties, spatial relations, and salience effects in REG.

Further, I will demonstrate that PRAGR can be used as a basis for enabling referential grounding dialogues by handling both REG and reference resolution (RR) using a single integrated reference handling mechanism based on the same underlying concepts. Therefore, although the focus of this thesis will be on the generation of referring expressions, I will also demonstrate the suitability of the presented approach for RR and its applicability in dialogic interaction.

1.2 Thesis Organisation

The rest of this thesis is organised as follows:

In **Chapter 2**, I will outline the scope of this thesis in detail. I will introduce the classic paradigm of REG and its extensions for handling various challenges for generating REs in realistic scenarios. From this starting point,

I will move towards a definition of reference which emphasises the nature of reference as collaborative action. Based on this understanding of reference, I will motivate and outline the contribution of this thesis with respect to the architecture of a dialogue system and the main challenges facing the REG community.

In **Chapter 3**, I will motivate and introduce the core contribution of this thesis, the **P**robabilistic **R**eference **A**nd **GR**ounding mechanism (PRAGR). Starting out from a discussion of what constitutes an optimal RE for a given situation, I reject the classical definition of the distinguishing description in order to develop an optimality definition which encompasses conceptual vagueness as an inherent constituent. I define the core concepts of PRAGR, Acceptability, Discriminatory Power, and Appropriateness and how they apply to REG and RR, respectively, and provide an illustrative evaluation of the core mechanism with respect to REG.

As a basis for further elaboration, **Chapter 4** contains a description of models for vague properties for use with PRAGR. After presenting the underlying approach of Conceptual Spaces (Gärdenfors, 2004b), I introduce all property models that will be used in the remainder of the thesis, covering the following domains: graded adjectives, colour, shape, projective terms, and projective regions.

Chapter 5 is concerned with extending PRAGR to handle some of the challenges for REG in realistic scenarios. The focus is on integrating spatial relations into REG where I discuss how the identifiability of the reference object can be factored into the evaluation of the optimality of complex descriptions. In this context, I also present an extension of PRAGR to handle salience. I further suggest a search algorithm for PRAGR which overcomes the complexity issues raised by combining vagueness and spatial relations in REG.

In **Chapter 6**, I present three empirical evaluation studies in order to demonstrate the usefulness of PRAGR for situated referential communication and its ability to handle the interaction of a variety of property models. I will present an evaluation of robot-robot interaction to evaluate the usefulness of vague properties for overcoming perceptual deviation, an evaluation of RR

where PRAGR interprets human produced utterances, and an evaluation by human subjects of REs generated by PRAGR.

In **Chapter 7**, I show how PRAGR can be integrated with a dialogue system to enable intelligent referential grounding dialogues. After a discussion of the ways in which PRAGR can support referential grounding dialogues, I present the integration of PRAGR with the DAISIE dialogue system framework and architecture for a simple referential grounding dialogue scenario.

Finally, in **Chapter 8**, I summarise the main findings of this thesis and discuss possible directions for future work.

Chapter 2

Scope of this Thesis

Reference is a fundamental aspect of communication which has been studied extensively from a wide range of disciplines, most notably philosophy, psycholinguistics and computational linguistics (van Deemter et al., 2012b). Accordingly, the generation of Referring Expressions (REs) is such a central aspect of Natural Language Generation (NLG) that almost all implemented NLG systems have some kind of Referring Expression Generation (REG) component (Mellish et al., 2006). In situated interaction, reference mainly has the purpose of identifying physical objects, for example as targets of manipulation or for anchoring motion instructions. In the example given in the introduction, Mary refers to a certain book, using the RE *the yellow book* with the intention that Amanda should bring her that particular book. When giving a route instruction in a city, one may say *turn left after the church*, thus referring to the church for anchoring the turn instruction.

But what exactly is reference? Due to its ubiquitousness, reference is a potentially wide and complex field, and although certain aspects of reference have been thoroughly studied, many limitations and restrictions still remain (Krahmer and van Deemter, 2012). As many researchers have pointed out, the concept and scope of reference are hard to pin down (Abbott, 2010; Krahmer and van Deemter, 2012; Searle, 1969). Therefore, researchers interested in modelling the breadth of human communication, such as Appelt and Kronfeld (Appelt, 1985; Appelt and Kronfeld, 1987), have de-

limited the phenomena differently from those interested in achieving useful results for applied NLG systems (Dale and Reiter, 1995; Reiter and Dale, 1992). Moreover, in the context of larger NLG systems, the delimitation of individual subtasks such as REG depends strongly on the domain and architecture of the overall system (Mellish et al., 2006).

In this chapter, I will give an overview of the range of the field and clarify the scope and focus of the reference problem as it will be addressed in this thesis. I will start by introducing the classic paradigm of REG and its extensions, questioning the philosophical underpinnings behind this paradigm, and proposing a collaborative view of reference as the foundation of the present thesis. I will then delimit the problem of reference as addressed in this thesis from the point of view of the overall architecture of a dialogue system, and from the perspective of concrete challenges to be solved within this thesis.

2.1 The Classic Paradigm of Referring Expression Generation

The most popular view of REG to date is based on the idea that the main challenge of REG lies in selecting, from a given set of properties of the intended referent, a subset of properties which uniquely distinguishes the target from all other objects, also termed distractors. Krahmer and van Deemter (2012, p. 203) summarise that "[a] substantial amount of REG research focuses on (...) 'first-mention' distinguishing descriptions consisting of a noun phrase starting with 'the' that serve to identify some target, and that do so without any further context."

In this paradigm (henceforth called classic REG, following van Deemter (2016, p. 82)), the task of REG is defined as follows: given a finite domain D with objects $d_1, d_2, \ldots, d_n$ with attributes $A = a_1, a_2, \ldots, a_n$ where each object is defined by a number of attribute-value pairs which are true of this object, find, for a given target object in D, a set of attribute-value pairs L whose conjunction is true of the target but not of any of the distractors. Such a set of attribute-value pairs is called a distinguishing description. Context,

in this paradigm, is modelled as "the set of entities that the hearer is currently assumed to be attending to" (Dale and Reiter, 1995, p. 236), while the paradigm abstracts away from the further physical, social and dialogic context of the utterance. In particular, this paradigm assumes that all relevant objects and their properties are part of the common ground between speaker and listener. Based on this paradigm, reference resolution (RR) should be a straightforward process of identifying the object for which all properties of an RE hold.

Not least due to these simplifying assumptions, classic REG has proven highly productive in stimulating a vast amount of research and useful algorithms starting with the Full Brevity Algorithm (FB) (Dale, 1989; Dale and Reiter, 1995; Reiter, 1990) which, following Grice's Conversational Maxim of Quantity: "do not make your contribution more informative than is required" (Grice, 1975, p. 45), aims at generating the shortest distinguishing description, and its two successors, the Greedy Heuristic Algorithm (GH) (Dale, 1989, 1992) and the Incremental Algorithm (IA) (Dale and Reiter, 1995; Pechmann, 1989). These three fundamental algorithms of classic REG will be discussed in detail in Section 5.2.1. In short, the GH provides a computationally more efficient approximation of the FB algorithm while the IA, justified by research on referential overspecification and attribute preferences in human subjects (Pechmann, 1989), incrementally selects properties according to a predefined preference order, thus frequently generating overspecified REs.

2.2 Extensions of the Classic Paradigm

In the decades following their publication, many extensions of these algorithms have been published, addressing a number of limitations of the original algorithms (for an overview, see Krahmer and van Deemter, 2012). Among others, extensions were proposed which can deal with spatial relations, graded properties, salience effects and referring to sets of objects. In the following, I will briefly discuss the challenges addressed by the community since the 1990s and their relevance to situated human-machine interaction,

while leaving the discussion of details to the respective chapters.

2.2.1 Spatial Relations

It has been shown that humans frequently use spatial relations for referring to physical objects, even in situations where they are not needed for achieving a distinguishing description (Viethen and Dale, 2008). Therefore, it is not surprising that the generation of REs including spatial relations is one of the most active areas of REG research. Dale and Haddock (1991) provide an early adaptation of the GH for relations, identifying a way to prevent infinite recursion using a constraint-based approach (see also Section 5.1.2). Kelleher and Kruijff (2006) present an adaptation of the IA to REG with relations which prefers absolutely discriminating relations over relatively discriminating relations. Krahmer and Theune (2002) present an extension of the IA for handling relations and salience. Krahmer et al. (2003) present a graph-based REG algorithm which covers relations and provides an elegant definition of distinguishing description in terms of subgraph isomorphism (see also Section 5.1.2).

However, contrary to the empirical evidence (Viethen and Dale, 2008), relational properties have often been treated as the least preferred properties, especially in extensions of the IA, where introducing relations poses the danger of *forced incrementality* – the unnecessary concatenation of large numbers of spatial relations in order to achieve discrimination (see also Section 5.2.2).

Moreover, most of the existing approaches offer only rudimentary treatment of the problem of reference object selection, which has shown itself to be a non-trivial issue in generating locative expressions (Barclay and Galton, 2008, 2013; Gapp, 1995a). Krahmer et al. (2003), for example, do not discuss the issue of reference object selection at all, treating relations as simply another way of achieving a distinguishing description. Kelleher and Kruijff (2006), on the other hand, restrict the set of possible landmarks at each step of the algorithm to those that, given a preliminary description, are not distractors of the target. The question of how different factors which have been

found to influence reference object selection in locative expressions, such as referentiality and salience of the reference object, discriminatory power and cognitive preference for certain expressions, can be weighted against each other has not so far been explicitly addressed in REG research.

2.2.2 Reference to Sets

Apart from referring to individual objects, efficient interaction also requires referring to sets of objects, either because several objects are relevant for a given task, or as a means to identify one object via its relation to a group. Several authors have proposed extensions of the classical algorithms for referring to sets (Gatt, 2007; Horacek, 2004; van Deemter, 2002). A crucial issue when referring to sets of objects is the question of which properties are shared by the objects in question while simultaneously serving to distinguish them from others. While van Deemter (2002) uses negation in order to increase the potential number of shared properties, Horacek (2004) proposes a *divide and conquer* approach of describing subsets of objects separately, if no adequate description for the whole set can be found. Bateman (1999) addresses the same issue by proposing the use of aggregation lattices which provide an overview of which properties serve to aggregate and/or discriminate certain sets of objects, to be combined with standard REG algorithms, although he does not specify the details of how this combination should be achieved.

2.2.3 Salience

Salience is equally important in situated interaction, as people do not perceive their environment with uniform attention. Rather, certain objects stick out perceptually or cognitively, they are more salient than the other objects (Clarke et al., 2013). Discourse context may also make certain referents more salient than others, for example a recently mentioned object will be more salient than one which has not been talked about at all in the current discourse (compare Krahmer and van Deemter, 2012, pp. 186–188). Empirical evidence indicates that salience phenomena influence both the production and

the resolution of REs. For example, salience effects influence the selection of properties and reference objects, and the amount of detail given in descriptions (Clarke et al., 2013; Hermann and Laucht, 1976; van der Sluis and Krahmer, 2004). Further, salience and focus frequently help listeners to disambiguate otherwise ambiguous REs (Clark et al., 1983; Kelleher et al., 2005; Strohner et al., 2000). Some extensions of the classic REG algorithms take salience effects into account by using salience to reduce the set of relevant distractors for a description, thus allowing for underspecified descriptions. For example, this would allow *the ball* as a description of an object rather than *the red ball* if the ball in question is the most salient ball in the visual and/or discourse context (compare Jordan, 2000; Krahmer and Theune, 2002; Passonneau, 1996).

2.2.4 Vagueness, Gradedness, Uncertainty

The assumption of crisp categories associated with classic REG causes difficulties when dealing with inherently vague concepts as seen with gradable adjectives: if a speaker says *the large mouse* in a context where several mice are present, this may not be a distinguishing description in the strict sense, as the property LARGE holds for all mice to a certain degree. A vast amount of research on categorisation starting with the early work of Berlin and Kay (1969) and Rosch (1973) has shown that effects of gradedness and vagueness are pervasive in categorisation, rather than being the exception, thus making it important to address these phenomena. Further, in situated interaction, there may be uncertainty regarding sensor information which means that a precise estimation of whether or not a certain property holds for an object may not always be available.

Phenomena of vagueness, gradedness and uncertainty have been addressed to some extent in the context of classic REG, as will be discussed in more detail in Section 3.1.4. van Deemter (2006), for example, provides an extension of the IA for handling graded adjectives by transforming numeric values to inequalities. The approach by Horacek (2005) takes into consideration term knowledge and perceptual and conceptual risks when considering properties

for REG, while Kelleher and Kruijff (2006) distinguish between relatively and absolutely discriminating relations in order to increase the likelihood of communicative success.

2.3 Changing Ideas about Reference

Although significant progress with respect to all of these challenges has been made (see Chapters 3 and 5), Krahmer and van Deemter (2012) conclude in their review of REG research that this has mainly been done in the form of isolated extensions of the existing classic algorithms, while the integration of different extensions into a unified approach remains unsolved. They state that "researchers have often zoomed in on one extension of the [Incremental Algorithm], developing a new version which lifts one particular limitation. Combining all the different extensions into one algorithm which is capable of, say, generating references to salient sets of objects, using negations and relations and possibly vague properties, is a non-trivial enterprise." (Krahmer and van Deemter, 2012, p. 190)

In order to investigate approaches to reference that may lead to an integrated solution for the stated problems, I will take a closer look at the philosophical underpinnings of classic REG in the following sections, and attempt to widen the horizon for tackling the issue from a new perspective.

2.3.1 Reference as an Intentional Act

The ideas underlying classic REG can be traced back to a philosophical discussion initiated by Frege (1892). Frege distinguished between the meaning of an expression (what he termed 'Sinn', or sense) and the expression's denotation ('Bedeutung' or reference). Meaning is the *mode of presentation*, and denotation is what is referred to (or a truth value for statements). Building on this distinction, Russell (1905) argues that in contrast to indefinite Noun Phrases (NPs) which merely presuppose the existence of the entity denoted, definite NPs additionally require that the conjunction of all properties ascribed to the entity in question is true for this entity, but not for any other

object. Thus, *Smith's murderer is insane* implies that there is one and only one murderer of Smith.

Donnellan (1966) questions Russell's view of denotation and argues that there are two different functions of definite noun phrases which cannot be determined independently of the context of their usage. On the one hand, there is the attributive function, for example saying *Smith's murderer is insane* when we do not know who murdered Smith, and want to make a statement about whoever fits the description (Donnellan, 1966). This attributive usage merely entails that there exists one and only one entity which fits the given description, while no assumptions are made with respect to the identity of that entity.

On the other hand, there is the referential function, for example saying *Smith's murderer is insane* when talking about Jones who is currently on trial for murder (Donnellan, 1966). In this situation, the goal of the speaker is to refer to Jones, independently of whether he is *in fact* the murderer of Smith or not (Donnellan, 1966). Therefore, in this context, the utterance carries the implication that there exists one and only one referent which carries the ascribed property of being Smith's murderer, namely Jones (Donnellan, 1966). However, according to Donnellan, that Jones is in fact Smith's murderer is only an implicature[1], not an entailment, as the reference can be successful and the utterance can be interpreted, even if Jones is in fact innocent (Donnellan, 1966).

Following this line of reasoning, reference is not an inherent property of some linguistic form, as both Frege (1892) and Russell (1905) assumed, but the intentional act of an agent with the goal of identifying a particular (Donnellan, 1966; Strawson, 1950). Or, in Searle's words, an RE is

> "[a]ny expression which *serves to identify* any thing, process, event, action, or any other kind of individual or particular (. . .). *Referring expressions point to particular things*; they answer the questions Who?, What?, Which?" (Searle, 1969, pp. 26–27, em-

[1]Donnellan uses the term *implication*, but it is clear that he means what Grice (1975) called Conventional Implicature – something that is suggested, but not entailed.

phasis: V.M.)

In this sense, an indefinite noun phrase can also be used referentially, for example I can say *Can you see a large green box?* when I have a particular large green box in mind that I want the listener to attend to.

Reference is therefore distinct from denotation, and thus truth. However, as Donnellan points out, the two are related. As a basis for communicative success, it is advisable to use descriptions which one considers true:

> "Because the purpose of using the description is to get the audience to pick out or think of the right thing or person, one would normally choose a description that he believes the thing or person fits. Normally a misdescription of that to which one wants to refer would mislead the audience. Hence, there is a presumption that the speaker believes *something* fits the description—namely, that to which he refers." (Donnellan, 1966, p. 291)

In classic REG, this goal of identification is formalised by modelling properties of objects as sets of boolean values and defining context as a set of distractor objects, abstracting from the physical, social and dialogic context of the utterance, and thereby restricting the task of REG to selecting those properties which discriminate the intended referent from all distractors, or in other words, selecting those properties which denote it unambiguously, as in Russell's 1905 analysis of definite descriptions.

2.3.2 Concepts as Condensed Experience

While this restriction of the task has proven highly productive (see Section 2.2), the question needs to be asked: what does it mean to say "x is the denotation of phrase 'C'" (Russell, 1905, p. 488)? Or, more to the point: what does it mean to say "[T]he proposition 'x is identical with C' is true." (Russell, 1905, p. 488)? The classic REG interpretation of these statements is based on the implicit objectivist assumption that the world comes readily separated into categories which need only be correctly identified by a human or artificial agent. According to this assumption,

> "[l]inguistic expressions get their meaning only via their capacity to correspond, or failure to correspond, to the real world or some possible world; that is, they are capable of referring correctly (say, in the case of noun phrases) or of being true or false (in the case of sentences)." (Lakoff, 1987, p. 167).

In this sense, a category is a set of entities in the real world which are naturally grouped together. The category DOG is "the set of all entities in the real world that are appropriately categorised as dogs" (Goldstone et al., 2012, p. 608). A concept, on the other hand, is a cognitive representation of a category or individual. Thus, the concept DOG is the psychological state which signifies thoughts of dogs (Goldstone et al., 2012, p. 608). For example, a book is either YELLOW or GREEN, and it can be determined *a priori* which of the two it is. If one then were to refer to a given object as *the green book*, this RE would refer correctly, if and only if the object in question were, in fact, both green and a book.

If we go back to the tragic death of Smith, this makes sense at first glance: Either Jones is the murderer of Smith, and thus a (the only, in fact) member of the set of entities in the real world who murdered Smith – in this case, the proposition *Jones is identical with 'Smith's murderer'* is true – or he isn't and the proposition is false. However, if things were that simple, courts would have less work, and criminal defence lawyers would not be charging so much money.

Humans have the ability to construe phenomena in different ways, by applying different concepts to them. With respect to Smith's death at Jones' hands, the same course of action may be conceptualised as MURDER or as an act of SELF-DEFENCE. In the example of the introduction, Amanda was able to construe the same book first as YELLOW, then as GREEN, due to the realisation of the mismatch between her original construal and that of Mary.

From an objectivist point of view, one may argue that Amanda was *wrong* about the colour of the book and corrected that mistake later. Amanda would be considered to have assigned the correct colour concept to the book if she had assigned the category that corresponds to the book's *real* colour in the world. But who is to decide what the *correct* borders of YELLOW and GREEN

are?

As Lakoff (1987) argues in his extensive discussion of objectivist semantics, this view is questionable. He argues for a position of **experiential realism**. According to this view, although there are certain objectively present features of the world – such as the way a certain object reflects light – the possibilities of categorising them are quite different for different creatures. Insects, for example, can usually see light in the UV spectrum, thus seeing colour differences (has UV component vs. does not have UV component) which humans do not see at all (Briscoe and Chittka, 2001). Moreover, research on insects shows that the number of different colour receptor cells, and the way the information provided by those cells is processed in the brain lead to different categorisation effects. For example, bees can discriminate two colours increasingly well, the larger the difference in wave length. Flies of the genus *Lucilia*, on the other hand, appear to have three distinct colour categories, and sameness or difference between stimuli depends entirely on whether the stimuli are in the same category or not, rather than their precise spectral difference (Briscoe and Chittka, 2001).

Such differences can also occur within the human species. For example, colour cones in the human eye have individually different distributions, even between humans who are not colour blind (Roorda and Williams, 1999). Beyond genetically induced differences, there may also be *perceptual deviation* based on situational or biographic differences. People standing at different positions may perceive colours, positions and other object features in a slightly different way due to lighting conditions, or imprecise estimation procedures (Spranger and Pauw, 2012). Also, different degrees of exposure to certain domains can influence sensitivity to perceptual differences in these domains (Goldstone et al., 2012, p. 621). All these kinds of perceptual deviation may influence both overall category structure and category assignment of individual instances (Spranger and Pauw, 2012).

Moreover, categorisation is more than just lumping together a set of things which are perceived as alike. It is lumping together a set of things which are perceived as alike in a way that is *relevant* to the agent, it is ultimately a way to make sense of the world (Steels, 2008). In our constant

interactions with the world, we pursue a range of different goals in ever-changing circumstances. In order to be able to achieve our goals, we group together aspects which are *meaningful* with respect to these goals in order to transfer our knowledge from one situation to another. We associate these aspects with a concept, a more or less stable mental state which we can consciously access (Dorffner, 1992) and represent externally, for example using symbols (Steels, 2008).

In this sense, concepts serve as equivalence classes —they enable us to abstract from the (superficial) differences of entities and treat them as alike for a given purpose (Goldstone et al., 2012). The grouping is performed according to some experiential *method* that allows us to identify whether any newly encountered object is part of the category we have formed (Goldstone et al., 2012; Steels, 2008). Steels (2008) defines the *method* of a concept as "a procedure to decide whether the concept applies or not" (Steels, 2008, p. 2). He further explains that "[t]he method could for example be a classifier, a perceptual/pattern recognition process that operates over sensorimotor data to decide whether the object 'fits' with the concept" (Steels, 2008, p. 2).

Links can be formed between different concepts based on personal experiences and available cognitive mechanisms which enrich concepts and enable us to infer adequate reactions based on the kind of objects we encounter. Concepts can be related to each other in many different ways, for example similarity or difference, hierarchical or causal relations, and many more (Steels, 2008). Seen from this perspective, in opposition to the objectivist definition of category given above, I propose the following definition of category in line with the perspective of experiential realism: A category is not a set of entities in the real world which is naturally grouped together, but rather a set of entities in the real world which is grouped together *by some agent or community of agents*.

Regarding the internal structure of concepts, a vast body of research from different disciplines such as computer science and cognitive linguistics shows that the concepts humans form have a complex and diverse internal structuring, as they are developed on the basis of complex and diverse experience (Dorffner, 1992; Frixione and Lieto, 2012; Gärdenfors, 2004b; Goldstone

et al., 2012; Lakoff, 1987; Rosch, 1973; Zadeh, 1965).

According to Gärdenfors (2004a, p. 18), concepts can be seen as convex regions in an n-dimensional conceptual space composed of one or more quality dimensions. The quality dimensions "correspond to the different ways stimuli can be judged similar or different" (Gärdenfors, 2011, p. 2). Some quality dimensions are integral: they are so closely related that they cannot exist independently. For example, an object cannot be assigned a value for hue without also assigning a brightness and saturation value. Other dimensions, such as size and hue, are separable. Gärdenfors (2011) defines a *domain* as "a set of integral dimensions that are separable from all other dimensions" (Gärdenfors, 2011, p. 2). He further distinguishes properties as a subtype of concept, which are convex regions within a single domain.

Gärdenfors (2004b) suggests modelling categories as prototypes, and determining category boundaries by performing a Voronoi tessellation such that each point in a conceptual space is considered to be a member of the category whose prototype is closest to it. However, there are many phenomena in human categorisation that are more complex. For example, sometimes categories are adapted to suit a specific context: when talking about human faces, we may use the full spectrum of colour terms while human faces only take on a limited subspace within colour space (Gärdenfors, 2004b). Gärdenfors (2004b) suggests that, in such cases, the colour space is restricted by the relevant contrast set (e.g. colours of human faces) such that the colours of the full space are mapped onto the restricted space. Thus, a RED or WHITE face have different colours from RED or WHITE wine or a RED or WHITE mug.

Similarly, the interpretation of graded properties such as LARGE and SMALL depends on the contrast set. Whether an object can be considered LARGE depends both on the type of object—a large mouse is usually smaller than a small elephant, but also on the distractor objects present in the situation—when looking at a table with many very large cups, one may consider a cup to be SMALL even if it is not particularly small for cups in general.

Further, some proposals have been put forward to extend Gärdenfors' (2004b) conceptual spaces approach to integrate vagueness, as some categories obviously lack clear boundaries (Douven et al., 2013; Mast and Wolter,

2013). In any particular context, there is no definite boundary of when a given cup is LARGE or SMALL. It is rather both LARGE and SMALL to a certain degree, and whether either term will be used in discourse may depend on a number of different aspects (Mast and Wolter, 2013; van Deemter, 2006).

Further, conceptual spaces are not necessarily nicely divided into mutually distinct concepts, but may have more complex structures. Some colour terms span fairly large areas of the colour space (e.g. RED) while others are very specific (e.g. CRIMSON or SCARLET). Although there are hierarchical relationships (e.g. CRIMSON is a kind of RED), these relationships are not always straightforward. For example, it is not clear whether MAROON is a kind of RED, or a kind of BROWN, or both.

Finally, as Lakoff (1987) convincingly argues, categories may be extended via metaphorical or metonymic processes, or based on family resemblances, yielding categories which would be hard to express in terms of conceptual spaces, as it would be hard to define them in terms of a limited number of quality dimensions.

While both the internal structure of these kinds of categories, and the relational links formed between them are to some degree determined by the inherent nature of the domain and the structure of the human sensorimotor system, the details depend strongly on individual experiences. Thus, ultimately, every person perceives, carves up and understands the world in a slightly different way, forming a rich individual semiotic network (Steels, 2008). What is YELLOW to one person may be GREEN to another. And this is where language comes in.

2.3.3 Language as a Coordinating Device

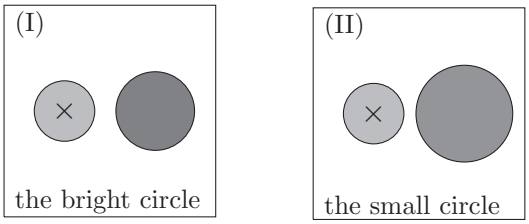
The usage of symbols is rooted in the desire to communicate information. In reference, we use symbols in order to draw the attention of the listener to an intended referent. From a semiotic point of view, reference means creating an external symbol which one believes will lead the listener to identify the intended referent, mediated via the *sense* – in the case of reference, the property concepts associated with the symbols. Whether Mary decides to

say *Amanda, bring me the yellow book.* or *Amanda, bring me the green book* involves an act of concept assignment where she chooses which concept, YELLOW_MARY or GREEN_MARY is more likely to lead Amanda to select this book in the given context. The chosen concept is then externalised via a symbol, *yellow*, or *green*. Reference resolution then means perceiving the symbol – *yellow* or *green*, retrieving one's associated concept – YELLOW_AMANDA or GREEN_AMANDA, and applying the experiential method of the concept to candidate referents in order to identify the intended referent.

Since the semiotic networks are individual, the meaning of symbols is also not universal. When Mary wants to communicate a concept and uses a symbol for it, Amanda will most probably associate the symbol with a similar concept, but never exactly the same concept. In order to achieve her communicative goal, Mary needs to choose her expression such that "the referent can be readily and uniquely inferred from the current common ground of speaker and addressees" (Clark and Bangerter, 2004). In other words, she needs to make a strategic decision about concept assignment, not only based on what she herself knows about the situation, but also based on her estimation of what Amanda knows about the situation, and her assumptions of what Amanda will make of her words.

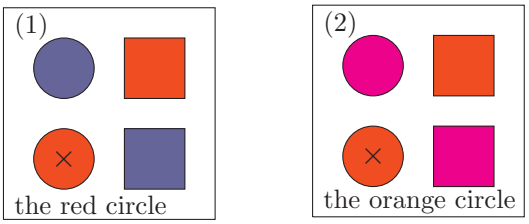
Thus, it is not only relevant whether calling the book *yellow* is discriminating according to some assumed objective colour assignment. Rather, it is important to judge how likely calling the book *yellow* will be discriminating *for the addressee*. Hermann and Laucht (1976) show that when two different graded attributes are available for identifying one object over another, the property with the largest object-distractor contrast is chosen. For an example, consider Figures 2.1a and 2.1b which both allow using either size or brightness for discriminating the target. In Figure 2.1a, the difference in brightness is perceptually larger than the difference in size, therefore brightness will most likely be chosen for referring to one of the objects. In Figure 2.1b, the size difference is perceptually larger, therefore size will be chosen.

But even within a single domain different conceptualisations may carry different risks of miscommunication. If there are other books which may be conceptualised as YELLOW_AMANDA, even if they are not YELLOW_MARY,



(a) (b)

Figure 2.1: Influence of context on property selection in REG (Mast et al., 2016). Situations where (a) brightness, and (b) size is the most salient property.



(a) (b)

Figure 2.2: Influence of context on linguistic encoding in REG (Mast et al., 2016). (a) Red is more distinguishing. (b) Orange is more distinguishing.

using the expression *yellow* yields a danger of miscommunication. This danger may be assessed by taking graded category membership into account, and thus making allowances for conceptual mismatch. Figure 2.2 shows how the same circle (marked with an X) may be called *the red circle* or *the orange circle*, depending on which other objects it occurs with. While it is debatable whether the circle is in fact RED or ORANGE, and whether the distractors in Figure 2.2b are RED or not, calling the referent *orange* in that scenario will improve the chance of successful communication.

On the other hand, specific knowledge about the addressee will also influence the choice of attribute, and concept assignment. If Mary knew from experience that Amanda had problems dealing with colour, she may decide that the colour of this book was not a good way to enable Amanda to identify it. In that case she may use a different property entirely. If, on the other hand, she had talked to Amanda about this book, or similar shades of

YELLOW_MARY before, she might know that Amanda was likely to see this book as GREEN_AMANDA so she may say *bring me the green book*, overriding her own preference for categorising the book as YELLOW_MARY. Ultimately, for Mary it is not relevant whether the book is *in fact* YELLOW_MARY or GREEN_MARY, but whether or not calling it *yellow* or *green* will make it more likely for Amanda to identify the correct book.

2.3.4 The Collaborative Nature of Reference

Up until this point, I have talked mainly about the production of REs, following the widespread separation between the production and comprehension of natural language – or, in computational terms, NLG and Natural Language Understanding (NLU). In one sense, this separation is appropriate, as generating natural language involves different kinds of processes than interpreting it. NLG is "the process of deliberately constructing a natural language text in order to meet specified communicative goals" (Dale, 1995). It requires making choices and pursuing goals. The central task of NLU, on the other hand, is disambiguating and inferring the goals of the speaker. The kind of problems encountered in NLG and NLU are therefore often different. The core problems in NLU are resolving ambiguity, and dealing with unknown or unexpected elements in the input (McTear, 2004, p. 91). In NLG, on the other hand, there exist problems such as selecting which information should be verbalized at all, or choosing between different ways of verbalising some given content (McTear, 2004, p. 99). As a consequence, NLG and NLU are usually dealt with separately, and are mostly realised as entirely separate components in applied dialogue systems (see Section 2.4.1). The same holds for the domain of reference, where much work has been done on either REG (for a comprehensive overview see Krahmer and van Deemter, 2012) or on reference resolution (RR) (see for example Funakoshi et al., 2012; Gorniak and Roy, 2005; Kelleher, 2006; Kruijff et al., 2006). Much less work treats both in a unified way (though see for example Appelt and Kronfeld, 1987; Kelleher and Costello, 2009; Kelleher et al., 2005; Zender et al., 2009).

However, as I have discussed above, referring always includes bridging a

conceptual gap, mediating via linguistic symbols between individual semiotic networks. As we have seen in the prior discussion, this process involves the danger of miscommunication. Therefore, it cannot be dealt with simply by a speaker unilaterally uttering an RE and a listener resolving that expression. Research has shown that reference is instead a collaborative process which requires agreement (Clark and Wilkes-Gibbs, 1986), and which forms part of a continuous process of grounding which seeks to ensure mutual comprehension (Garrod and Anderson, 1987). For example, consider the following dialogue in a matching task where a director has to convey the order of a set of tangram figures to a matcher (Clark and Wilkes-Gibbs, 1986, p. 22, formatting: V.M.):

> [Director:] Urn, third one is the guy reading with, holding his book to the left.
> [Matcher:] Okay, kind of standing up?
> [Director:] Yeah.
> [Matcher:] Okay.

In this example, both participants collaborate in a joint effort to achieve understanding. The director tries a first RE, which the matcher expands in an attempt to confirm their interpretation of the expression. The dialogue ends with both participants verbalising their confidence that successful reference has been achieved. In the example presented in the introduction to this thesis, Mary and Amanda have a short dialogue about fetching a book. In this dialogue, Mary and Amanda also collaborate on achieving grounding, though in a slightly more complex way: Mary first provides an RE, *the yellow book on my desk*, and Amanda provides an RE of her own, *the one in front of the coffee cup*, in order to confirm that they are talking about the same object—similar to the expansion in the example above. Mary detects the miscommunication, and provides an overt correction, *not the green one, the yellow one*, which allows Amanda to adapt her construal and identify the originally intended referent. As in the example above, Amanda confirms that she considers the reference dialogue to have reached a successful ending.

This kind of collaboration is necessary, because common ground is not

a well-established homogeneous body of mutually known facts. Some elements may be firmly established, others in doubt, and yet others may not have been assessed yet as to whether they are part of common ground (Clark and Bangerter, 2004). Common ground can range from knowledge to suppositions, and can be based in perception, abstract inference, or any other source of information available (Clark and Bangerter, 2004). For example, from prior discourse, Mary may have gotten the impression that Amanda's colour perception was less than perfect, but she may not be sure how much it will interfere with their interaction. Or she may know for certain that a given shade of GREEN_MARY corresponds to YELLOW_AMANDA due to some prior interaction involving that hue.

In order for communication to be successful, interactants need to continuously build and reassess their common ground, a process called *grounding*, thus bridging gaps between individual views of the world to reach a mutual understanding. Grounding is performed via a range of mechanisms from automatic priming to implicit or explicit negotiation (see for example Clark and Wilkes-Gibbs, 1986; Garrod and Anderson, 1987; Pickering and Garrod, 2004). If Amanda initially conceptualises the book as YELLOW and then adapts her conceptualisation to GREEN based on the feedback from Mary, she is not in fact correcting a mistake, but rather participating in a collaborative process of grounding. In this understanding, language is a historically evolved conventional coordination device for a recurrent coordination problem, namely that of referring to entities and concepts (Clark and Bangerter, 2004; Steels, 2008, p. 31).

On a larger scale, the semiotic networks within a group get progressively coordinated via constant collaborative grounding efforts between individuals (Clark and Bangerter, 2004; Steels, 2008). Thus, language as a cultural tool provides an anchor to synchronise the categorisation processes of individuals in a linguistic community and gives us a way to reach an understanding that we are talking about the same thing. In a study on collaborative maze-games played in pairs, Garrod and Doherty (1994) show that linguistic and conceptual convergence occurs not only between individual speakers, but also across a (simulated) community of speakers over the course of several pairwise

interactions. In experiments on language evolution in robot communities, Bleys et al. (2009) show that enabling adaptation of grounded categories increased communicative success within a robot community over a sequence of interactions.

Coming back to the question of treating REG and RR as separate problems, the collaborative nature of reference indicates a strong link between the two. Estimated understandability of an RE by the listener plays an important role in generating REs, and insecurities or ambiguities in RR may be overcome by generating a clarifying RE of one's own. A more detailed discussion of the relationship between REG and RR will be given in Section 3.1.1. For the purpose of the current argument, it suffices to conclude that successful grounding is only possible if there are at least some common structures underlying RR and REG which ensure that the dialogic grounding process can be co-ordinated throughout role switches of the interactants.

2.3.5 The Case for Referential Grounding in Situated Interaction

Based on the argument I proposed in this section, we can conclude that calling a book *yellow* or *green* is to some degree a matter of strategic choice, restricted by the physical qualities of the world (e.g., the way the object reflects light, which other objects are present in the scene), our own perceptual system (e.g., the way we can perceive light), the linguistic norms of our community (e.g., the way the spectrum of light waves is normally classified in that community), our discourse history with the current interaction partner (e.g., how expressions and concepts have been grounded with this partner), and our specific goals in the current situation (e.g., referring to an object vs. describing it). Further, I have argued that such a decision is always a momentary, strategic decision to construe an object or property in a certain way, and not the ultimate decision about the truth of certain propositions. Thus it is subject to processes of negotiation and adaptation.

If we look at reference from this perspective, it becomes clear that some meta-knowledge about concept assignment needs to be available to any agent

when producing and resolving referring expressions in order to achieve successful communication. But in comparison to human-human interaction, an artificial agent communicating with a human is at a large disadvantage. Humans within a linguistic community have relatively well-synchronised semiotic networks due to their similar bodies with similar sensorimotor systems, due to their similar experiences and a long common cultural heritage of continuous grounding. A situated artificial agent however, for example a robot, is a being of an entirely different nature. Its sensory capacities are built and structured differently from humans', and while it is possible to emulate relevant aspects of a human's semiotic network in an artificial agent to some degree, the way humans perceive and understand the world is not well enough understood to come even close to matching it. For example, while a robot perceives colour in terms of the wavelength of the light an object reflects, humans subconsciously perform a whole range of abstractions and transformations based on the setup of the sensory system, and life experiences, which are not sufficiently understood to be adequately modelled. Imagine sitting in a meeting room with a WHITE wall on which a projector projects some slides with WHITE text on a BLACK background. To a robot, the BLACK background of the projected image would seem to be the same colour as the WHITE wall surrounding the projection, as it has exactly the same values of hue, lightness and saturation (for further examples, see Lotto and Purves, 1999).

Due to the higher risk of conceptual mismatch in situated human-machine interaction, the ability to anticipate this mismatch and to deal with it once it occurs is particularly important for situated agents. Based on these considerations, the use of grounded representations (Roy, 2002) which allow bottom-up sub-symbolic perceptual features to influence symbolic processing provides a promising new approach for reference handling, particularly in the domain of situated interaction.

This endeavour goes beyond REG or RR alone, as it requires a component which provides a common basis for the processes of understanding and generating REs. In situated interaction, access to the underlying basis for categorisation is a great advantage, as it enables reaction to, and dialogue

about mismatched models of the world in a much more subtle way than a system based on crisp properties. For example, if the user commands the robot to *go to the large green box*, a robot sensitive to the information underlying categorisation may be able to identify the correct referent, even if it would rather call this object *yellow* if it were to produce an expression of its own. Further, if the sensory information warrants *green* as an acceptable categorisation, this mismatch might be corrected without any further negotiation having the robot respond with *okay, I'll go to the large green box.* While, if the mismatch were fairly large, the robot might react with a clarification question using additional features which, in the given situation, promise more certain categorisation: *do you mean the large box to the left of the red ball?*

Based on these assumptions, the central goal of this thesis is to provide a computational mechanism for reference in situated interaction which integrates meta-knowledge about concept assignment into the process of generating and resolving REs, and which is therefore capable of strategically using flexible concept assignment in order to achieve communicative goals. As the system should be able to engage in referential grounding dialogues, my aim is to provide an integrated mechanism which performs generation and resolution of REs based on the same underlying representations, and using the same fundamental mechanism for achieving both tasks. In the following section, I will delimit the goal and scope of this thesis in more detail.

2.4 Contribution of this Thesis

While the previous section has dealt with the nature of reference on a more general level, in this section I will delimit the aspects of reference which will be addressed in this thesis. First I will determine the scope of the proposed system within the general architecture of a dialogue system. I will then identify the particular challenges of REG which will be dealt with in this thesis.

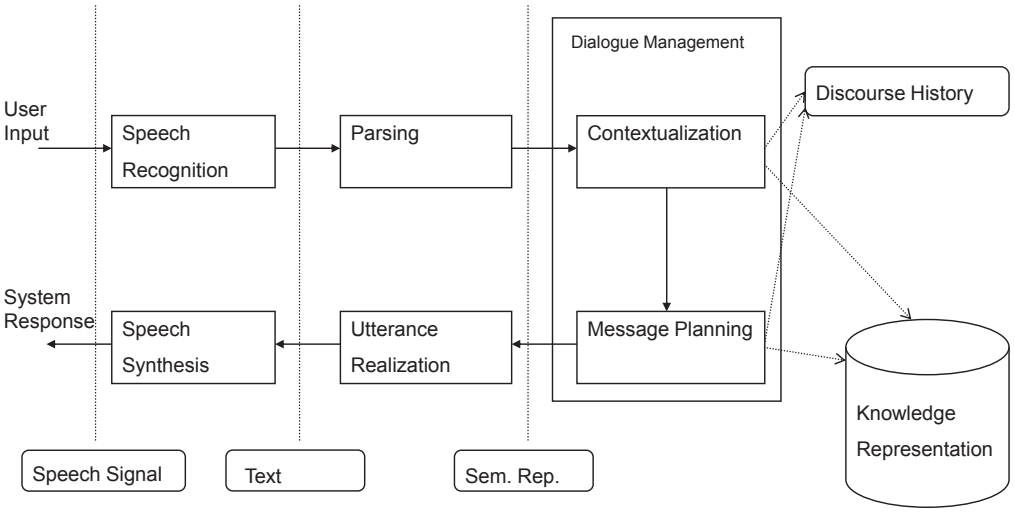


Figure 2.3: Typical architecture of a dialogue system.

2.4.1 Reference in the Architecture of a Dialogue System

In order to fully understand the scope of this work, and its role in a dialogue system for situated interaction, I will now briefly discuss the typical architecture of a dialogue system, and the roles REG or RR modules usually play in such a system. I will then proceed to outline the particular tasks that the system presented in this thesis performs, the kind of input it requires and the kind of output it produces, and which demands this places on dialogue management and knowledge representation.

While there is much diversity with respect to overall architectures of dialogue systems depending on a number of factors, there is some level of agreement on the prototypical structure of a spoken dialogue system (McTear, 2004; Ross, 2009). Figure 2.3 shows a typical architecture of a spoken dialogue system.

A typical dialogue system has an NLG and a Natural Language Understanding (NLU) component which are tied together via a dialogue manager. The outer NLG and NLU components each perform a mapping between different types of representations. On the NLU side, speech recognition maps speech signals to textual representations, and parsing maps textual repres-

entations to semantic representations. On the NLG side, surface realisation maps semantic representations to text, and speech synthesis maps text to speech signals. As Ross (2009) shows in an analysis of several advanced dialogue systems, the dialogue manager component itself usually also has distinct modules for integrating user dialogue moves (contextualisation) and planning the next system move (message planning). According to Reiter (1994), the message planning module may be further divided into a content selection, and a sentence planning module. Overall, the typical dialogue system follows a standard *pipeline* architecture, where the output of one processing module is used as the input for the subsequent module.

While Reiter (1994) argues that the NLG side of this architecture is a *consensus architecture*, in a more recent study of 20 complete NLG systems, Mellish et al. (2006) show that this consensus architecture is in fact an idealisation. They analyse seven low-level generation tasks and show that almost all tasks can occur in either of the three main modules, often spanning the range of two modules even within one system. Thus, while for example the rhetorical planner occurs almost exclusively in the content determination module, REG components can be found in content determination, sentence planning, and surface realisation modules, though most often they are found in the sentence planning modules.

It is not surprising that REG may play different roles in applied NLG systems, depending on the goals of the system and the overall architecture. Content selection is a highly relevant aspect of REG in situated interaction where objects with different properties need to be identified while this is less relevant for generating texts where most referents are assumed to be known. When generating longer texts, REG is mainly a part of sentence planning, as the division of content into separate sentences requires anaphoric expressions such as *she*, which refer to entities mentioned prior in the discourse. Sentence planning may also be relevant for the generation of complex referring expressions involving spatial relations, as the availability of a reasonable way of expressing certain content may also influence whether this content should be selected for expression at all (Horacek, 1997). Further, the discussion on the nature of reference (see Section 2.3) has shown that there is also an

interaction between concept assignment and property selection for REG.

While a large amount of REG research, particularly classic REG, focuses on content selection – selecting the properties of the target that should be included in the description (Krahmer and van Deemter, 2012), some researchers have gone beyond the pipeline approach, exploring REG systems with a tighter coupling of content selection to other levels of generation, in particular surface realisation (Horacek, 1997; Stone, 1998). Some work has also been done concerning the interaction of content selection with concept assignment (Horacek, 1997; Roy, 2002; Spranger and Pauw, 2012), an issue which will be addressed in detail in Chapter 3.

Reference resolution, on the other hand, has mainly been equated with coreference resolution: the identification of different REs which refer to the same discourse entity (Zhekova, 2013), usually by relying on surface or syntactic features such as word distance (Tetreault and Allen, 2004) (compare also Jurafsky and Martin, 2008, Chapter 18), with little work done on incorporating semantic information (for an example, see Tetreault and Allen, 2004). In particular, most work focuses on identifying discourse referents while ignoring the issue of relating expressions to entities in the real world, or some model thereof. In the dialogue architecture presented above, both tasks fall under the heading of *contextualisation*, possibly reflecting the lack of explicit treatment of the latter. Linguistically speaking, this existing body of work is mostly concerned with *endophoric* REs – REs which point towards other REs within the text. The lack of interest in mapping descriptive expressions to the corresponding entities in a situation model is understandable if one operates under the assumptions of objectivist, crisp categories, as classic REG does. In that case one can assume that either a distinguishing description can be found, and will be uttered – in which case recovering the intended referent is trivial, or no such description can be found – in which case, trying to identify the intended referent is futile. However, if one assumes the perspective of language as a coordinating device which mediates between individual semiotic networks, and which needs to navigate perceptual deviation, the problem of RR based on *exophoric* REs – REs which point to entities in the environment rather than within the text itself, such

as many definite descriptions – becomes non-trivial and part of a joint dialogic effort. Some work on identifying referents in a situation model has been done by Kelleher et al. (2005), who provide an RR mechanism which resolves ambiguity of exophoric references by relying on visual salience, as humans do (Clark et al., 1983). Further, Kelleher (2006) integrates graded category membership into RR, although he does not discuss this issue explicitly.

The focus of this thesis lies on the interaction between concept assignment and content selection on the side of NLG, and on the interaction between concept assignment and contextualisation proper in NLU. Thus, in terms of the typical architecture of dialogue systems presented above, this thesis addresses only aspects of dialogue management, operating on semantic representations. Issues relating to surface forms will not be discussed. Rather, it is assumed that the component presented here generates a semantic representation of the entities, properties and relations to be mentioned, and leaves issues of surface realisation, such as pronominalisation or sentence planning to separate modules. Likewise, in RR, coreference resolution is not treated, but rather it is assumed that the RR component receives input enriched by a separate coreference resolution module which accumulates information from different turns. Take, for example, the following hypothetical interaction:

(1) Human: *Go to the red box.*
(2) System: *Do you mean the one on the large table?*
(3) Human: *No, I mean the one on the floor.*
(4) System: *Okay.*

In this example, given the identity of the assumed referent, the output of the REG component for generating an RE for utterance (2) would be [RED(1), BOX(1), SUPPORT(1),(2), TABLE(2), LARGE(2)]. This may be realised as either *the red box which is on the large table*, *the red box on the large table*, or *the one on the large table*.

The input of the RR system in order to interpret utterance (3) would need to be [RED(1), BOX(1), SUPPORT(1),(2), FLOOR(2)]. The output would be the most likely referent for this combination of features. Thus, the fact that (1) and (3) refer to the same entity needs to be established before accessing

the reference handler, based on the dialogue structure and the resolution of the one-anaphora.

A processing level which is usually not considered in work on REG or RR is knowledge representation. In contrast to most prior work, this thesis also incorporates aspects of knowledge representation. The core goal of this thesis is enabling flexible use of concepts in order to increase communicative success. Therefore, the proposed system – unlike classic REG – requires probabilistic information on the acceptability of relevant categories for all objects contained in the situation model. The crucial point here is to retain the modular architecture which separates knowledge representation from contextualisation and content selection, while at the same time ensuring that the information relevant for decision-making is available to the respective components. Thus, the knowledge base is required to provide a probabilistic value of acceptability for each object-category pair.

The core mechanism for REG and RR then *uses* this kind of information in order to generate semantic specifications for REs, or identify potential target objects. Although the probabilistic modelling of particular domains is not the primary goal of this thesis, in order to demonstrate the usefulness of this approach – particularly regarding the handling of spatial relations in reference, this thesis also provides a number of property models which provide the required probabilistic values. This is achieved by using a modular ontological representation which provides probabilistic mapping procedures from sensory data to qualitative concepts that enable an estimation of graded category membership for each object-category pair, comparable to the *methods* for symbol grounding discussed by Steels (2008, p. 2). In this thesis, probabilistic models of projective terms, colour, and a number of graded properties such as volume are provided, and the overall system is evaluated based on those models.

Finally, while the proposed system does not include dialogue management proper, i.e., decision making about dialogue moves or non-verbal actions, it does provide information directly relevant for decision-making. The prototypical dialogue system architecture presented above has the form of a pipeline: each module is assumed to take the input of the prior module, and

provide output that will be used by the following module. Once a decision for a certain dialogue move has been made, it is expected to be realised by going first through message planning, then realisation, then speech synthesis. In principle, the proposed system may be used in this way. However, it may be beneficial for the decision-making process to know how good a proposed RE is, or how certain the assignment of a potential referent to a given expression is. For example, if the RR component returns a potential referent, and provides the information that this referent is most likely the intended referent, and the potential for miscommunication is low, a system may proceed directly to confirming a request and acting upon it. If, on the other hand, the potential for miscommunication is high, the system may initiate a grounding dialogue. Likewise, if the best RE has a low appropriateness value, the system may select to first ground potential reference objects, in order to then proceed to talk about its intended referent, rather than mentioning it directly.

Finally, the proposed system is implemented in a working prototype using the DAISIE dialogue system (Ross, 2009) in which the presented module is integrated. While the proposed system is implemented in such a fashion that it may be easily integrated into existing dialogue systems, the actual integration is not my focus, and therefore does not constitute a part of this thesis.

In summary, the thesis presented here covers the modules of content selection on the side of NLG, and contextualisation on the side of NLU, while ignoring surface realisation, and coreference resolution. Further, I propose a probabilistic interface between knowledge representation and reference handling, and demonstrate a number of probabilistic knowledge representation modules which implement this interface. Finally, the presented system provides evaluative information about generated REs and resolved referents which may be used by a dialogue manager. The system presented is implemented in a way that allows its ready integration into existing dialogue systems, as will be discussed in Chapter 7.

2.4.2 Challenges

After delimiting the scope of this thesis from the point of view of the architecture of a dialogue system, I will now proceed to delimit the scope from the point of view of the challenges the REG community is currently addressing.

As I have briefly discussed in Section 2.2, the REG community has dealt with a number of challenges in the last decades, mainly in the form of extensions of the classic algorithms proposed by Dale and Reiter (1995). As Krahmer and van Deemter (2012) note, most of these extensions deal with one or more challenges in often very limited scenarios. The contribution of this thesis is therefore focused on providing a single mechanism that provides an integrated solution for several of the challenges that have been addressed during the last decades, namely: using spatial relations in REG; addressing the problems of vagueness, gradedness and uncertainty; and integrating salience measures into the decision process in both REG and RR.

Krahmer and van Deemter (2012) argue that the integration of REG approaches with standard knowledge representation frameworks is the key to overcoming these limitations, and discuss the limitations and advantages of different approaches. While using those approaches has enabled REG researchers to successfully deal with some of the challenges listed above, some fundamental problems remain, as will be discussed in Chapter 3. In this thesis, I take up the suggestion to use an established computational framework for dealing with the task of REG, but rather than knowledge representation frameworks, I rely on Bayesian statistics as the foundation for the REG mechanism presented.

Going beyond the presentation of an integrated solution for well-explored challenges, I intend to provide an approach to reference which allows dealing with all of these challenges from the perspective of probabilistic representations, thus addressing the danger of conceptual mismatch, and providing the foundation for artificial agents to engage in interactive grounding dialogues.

For all of the particular challenges covered, I will further address how this integrated approach deals with some fundamental limitations of prior solutions, focusing on spatial relations and the challenge of dealing with

vagueness, gradedness, and uncertainty. While a detailed discussion of existing work and open questions regarding the individual challenges will be performed in Chapters 3 and 5, at this point I will briefly outline which specific aspects of the individual challenges I address in this thesis.

Vagueness, Gradedness, Uncertainty

Unlike prior work where little attention has been paid to aspects of vagueness, gradedness and uncertainty, these issues form the central challenge of this work. As will be discussed extensively in Section 3.1, I intend to demonstrate with this thesis how substituting binary truth values with probabilistic information about category membership not only allows us to generate and resolve REs with graded properties such as LARGE vs. SMALL, but also allows us to take into consideration the potential for conceptual mismatch stemming from insecure sensory information, and conceptual mismatch between communication partners. Thus, contrary to approaches such as those presented by van Deemter (2006), the present work does not treat graded category membership as an exceptional case that needs to be integrated into existing frameworks by making special additions – which yield their own problems and hinder integration with other singular extensions, but rather as the norm that underlies the fundamental processes of collaborative reference, thus allowing the flexibility necessary to bridge the gap between the semiotic networks established by different individuals' unique life experiences.

Spatial Relations

As discussed in Section 2.2, spatial relations in REG are often treated as least preferred properties, a fact which runs counter to empirical evidence. In this work, I will show how the proposed Bayesian mechanism for reference handling is capable of balancing preferences for spatial relations in a way that their higher complexity is considered in the REG process alongside their potential for discrimination of the intended referent.

Further, I will show how by integrating salience and graded category membership, a relevant number of important criteria for reference object

selection are taken into consideration when generating REs with spatial relations, namely reference object salience, search space reduction, presence of distractor objects, and referentiality (compare Barclay and Galton, 2013; Gapp, 1996). Moreover, the suitability of the reference object has an impact on the question whether a spatial relation will be preferred to an alternative non-relational description of the target object.

Salience

Regarding salience, the major contribution of this thesis is to show how object salience can be included in the presented reference handling mechanism, and thus improve reference object selection processes. Further, I will discuss how the presented approach may be expanded in order to model the influence of object salience on the length of expressions.

While it would be interesting to investigate how the probabilistic approach to reference presented in this thesis would be suited to issues of reference to sets and negation, I will leave those issues for future work.

2.5 Summary

In this chapter, I have given an overview of the field of reference and motivated the approach taken to reference in this thesis. Questioning the philosophical underpinnings behind classic REG, I have explained how focusing on the collaborative nature of reference, and the goal-oriented, flexible nature of conceptualisation may aid in developing a reference handling component which allows for smooth and natural situated human-machine interaction. I have further delimited the contribution of this thesis, indicating as the major contribution the development of a probabilistic mechanism for REG which can integrate gradedness, spatial relations, and salience into one coherent mechanism for both REG and RR, thus providing a straightforward unified solution to problems which have so far been considered mostly as separate issues.

Chapter 3

A Probabilistic Reference and Grounding Mechanism

In the previous chapter, I have argued for the necessity of a reference handling mechanism which allows an artificial system to engage in referential grounding dialogues with humans, in the face of vague categories and perceptual and conceptual mismatch. In this chapter, I will motivate and present the core mechanism of this thesis, the **P**robabilistic **R**eference **A**nd **GR**ounding mechanism (PRAGR). In Section 3.1, I will discuss different approaches to the concept of *optimality* in REG, focusing on the integration of vagueness into REG, and explain in detail the approach taken to optimality within this thesis. Following this discussion, in Section 3.2, I will present in detail the first major contribution of this work, the core mechanism and probabilistic semantics of PRAGR. In Section 3.3, I will then present an example-based evaluation of the basic PRAGR mechanism.

3.1 The Optimal Referring Expression

In each given situation a whole range of linguistic expressions could be used to refer to a certain target object. Even when focusing on the standard paradigm of first mention distinguishing descriptions, as it has been defined in Section 2.1, a number of potential descriptions for each object remain.

Figure 3.1: Example of a scene where several descriptions could be used to refer to an object [1].

For example, in Figure 3.1 both *the red ball* and *the small red ball* would be acceptable descriptions for the same object.

Gatt (2007, p. 32) distinguishes two central problems that any mechanism for generating REs must solve: Firstly, the mechanism needs some function that determines whether one referring expression is preferable to another. This *optimality definition* constitutes an ordering over all possible descriptions, with the optimal description at the top of the ordering (Viethen, 2011, p. 46). Secondly, it requires a procedure for finding the optimal description (or a sufficiently close approximation) in the search space (Gatt, 2007). This procedure constitutes a search algorithm that determines in which order potential descriptions should be evaluated. Viethen (2011) emphasises the central role of the optimality definition, as it "shapes every step of an algorithm as it chooses between preliminary descriptions" (Viethen, 2011, p. 45). Depending on the nature of the optimality definition, different search strategies may be possible or desirable. Thus, in this chapter I will first introduce the optimality definition proposed in this thesis, while possible solutions for the search problem will be addressed in Section 5.2. There, I will also discuss the implications of the proposed optimality definition for the search problem.

[1] Image created using the POV-Ray 3D rendering software http://www.povray.org/

As REG is the primary focus of this work, this chapter will focus mainly on the aspect of REG. Where appropriate, I will mention relevant considerations regarding RR. Moreover, I will provide a brief description of the most straightforward transfer of the PRAGR mechanism to RR.

In the following, I will discuss different concepts of optimality as a basis for defining the approach to *optimality* taken in this thesis. First, I will discuss different perspectives on optimality, considering firstly, the dimension of human-likeness vs. understandability, and secondly the role of data in the design cycle of an REG system. From this broad view, I will proceed into an analysis of several concrete *optimality criteria* that have been proposed in the literature.

3.1.1 Human-likeness vs. Understandability

In the 1980s, the goal of researchers such as Appelt and Kronfeld (Appelt, 1985; Appelt and Kronfeld, 1987; Kronfeld, 1990) was to model human language production and understanding, with a focus on testing the boundaries of conventional accounts by integrating difficult and less frequent cases.

In the early 1990s, following the lead of Reiter and Dale (Dale, 1989, 1992; Dale and Reiter, 1995; Reiter and Dale, 1992), the field began to focus more on a restricted area of research with the primary goal of producing useful descriptions. Thus, Dale and Reiter argue for *understandability* as the central criterion of optimality from the perspective of pragmatics: if an interaction is to be successful, it is necessary to produce REs which the listener can understand as easily as possible (Dale, 1989, 1992; Dale and Reiter, 1995; Reiter and Dale, 1992). In this case, the goal is to enable the user to understand an RE with minimal cognitive effort. In the words of Dale and Reiter (1995, p. 6), identifying the correct target object "should not require a large perceptual or cognitive effort on the hearer's part".

Criticism of the unnaturalness of the descriptions produced by these early algorithms, and the increasing availability of dedicated REG corpora then brought a turn back towards viewing optimality from a perspective of *human-likeness*. Thus, a good RE was increasingly considered to be one which is as

similar as possible to REs produced by humans in a comparable situation. This development built to a large extent on the Incremental Algorithm (IA), an algorithm based on the assumedly inherent preferences of humans for certain attributes. With the availability of semantically transparent corpora of human-produced referring expressions such as the Drawer (Viethen and Dale, 2006), GRE3D3 (Dale and Viethen, 2009) or the TUNA corpus (Gatt et al., 2008, 2009), REG research experienced an "empirical turn" along with the discipline of computational linguistics as a whole (van Deemter, 2016, p. 105ff). Beyond simple evaluation, this development has made it possible to use Machine Learning methods to optimise REG systems to directly match human-produced REs (e.g. Bohnet, 2008; Jordan and Walker, 2000, 2005; Stoia et al., 2006; Viethen, 2011).

At first glance, one may assume that human-likeness and understandability are merely two separate goals serving different purposes. If one is concerned with modelling human cognition, human-likeness is an obvious goal, as a model that behaves similarly to humans is more likely to represent the processes leading to human decisions than one that does not create human-like output. On the other hand, if one wants to produce a useful system that can successfully communicate with humans, understandability may be the more reasonable criterion of optimality.

This distinction also holds for the seemingly more straightforward case of RR. Although there is only one intended referent, one should not automatically assume that resolving an expression to that referent is the goal of RR. If one wants to model human cognition and behaviour, the best RR may be the one that reproduces resolution errors typically made by humans.

However, beyond the distinction of cognitive modelling versus applied systems, there are complex interactions between human-likeness and understandability. On the one hand, humans adapt their behaviour to make understanding easier for the listener (Krauss and Weinheimer, 1966; Wilkes-Gibbs and Clark, 1992), as is made explicit in game theoretic approaches to REG (e.g. Golland et al., 2010). Thus expressions which are similar to humans' utterances are more likely to be easy to understand. Moreover, in RR, listeners have been shown to take into account the decision-making process of the

speaker in order to reduce ambiguity (Frank and Goodman, 2014).

The strong interrelation between human-likeness and understandability also means that there is some degree of overlap between the two perspectives with respect to incorporating specific evaluative measures. One can argue for considering salience effects in REG both from the perspective of making REs easier to understand and from the perspective of modelling cognitive effects which underlie a speaker's choices (compare Clarke et al., 2013; Kelleher and Costello, 2005).

On the other hand, it is also clear that humans have certain perceptual and cognitive limitations which are at work when producing language, thus speakers may not always be able to make choices which are optimal for the hearer. Haywood et al. (2005) show that speakers balance ease of production with considerations for ease of comprehension. In a study on object reference by Horton and Keysar (1996), speakers took less consideration of listener's requirements when they were under time pressure than otherwise. This means that, despite the interactions mentioned above, we should expect a certain degree of divergence between REs that are maximally similar to those generated by humans, and REs which are maximally efficient for the purpose of communication. This has been confirmed by the TUNA corpus evaluation studies which evaluated several systems according to human-likeness, human quality judgments and task success (Belz and Gatt, 2008; Gatt and Belz, 2008; Gatt et al., 2008, 2009). Further evaluation of the results of these studies showed no significant correlation between human-likeness and task success (Belz and Gatt, 2008; Gatt et al., 2009), leading the authors to conclude that "a system's ability to produce human-like outputs, may be completely unrelated to its effect on human task-performance" (Belz and Gatt, 2008, p. 200).

Further, Viethen (2011) notes that behaviour regarding REs has been shown to differ substantially between humans, and even within humans depending on circumstances (Viethen, 2011, p. 104). Thus, it is not even clear what one should model if one aims to model human behaviour. A number of systems take inter-subject variation into account by incorporating subject-specific preferences into REG systems (Bohnet, 2008, 2009; Di Fabbrizio

et al., 2008). Gatt et al. (2013), on the other hand, address this issue by creating a non-deterministic system for REG which varies between equally discriminating descriptions according to relative preferences – using probabilistic selection between equally discriminating properties, or even adding redundancy with a likelihood based on empirical evidence.

A different approach that one can take here is using understandability as an optimality criterion in cognitive modelling in order to provide a model of an idealised human. If a system equipped with such a mechanism is capable of successfully communicating with humans or other artificial agents, it can be assumed that a relevant aspect of the underlying mechanism has been modelled, even if other aspects, such as cognitive limitations of the speaker, have not been addressed and thus human-likeness in the strict sense is not achieved. There are a number of advantages of using understandability as the basis for an idealised model of human referential behaviour. Firstly, it provides a transparent decision mechanism which can then be enriched by integrating further influencing factors (such as salience effects – see Section 5.1.4, or limited processing capacity). Secondly, it allows the use of a single underlying mechanism for dealing with both REG and RR, while aiming for human-likeness in both may require using fundamentally different mechanisms. Thirdly, by aiming towards effective communication, the likelihood of successful application in real human-machine dialogues is increased, thus allowing for the examination of more complex behaviours, such as referential grounding.

Based on these considerations, the primary goal of this work is to present a system that can generate REs which are effective in situated human-machine interaction, and which may serve as an idealised model of what humans are *trying to achieve* in referential communication. Thus, the main perspective adopted in this thesis is that of communicative success, i.e. understandability in REG, and correct identification in RR. Due to the discussed interactions between human-likeness and task success, research on the production and comprehension of REs are considered together rather than separately, in order to gain a thorough understanding of the processes of REG and RR as a basis for the design of the proposed PRAGR mechanism.

3.1.2 The Role of Data in REG

The role of empirical data in the definition of optimality is another dimension in which approaches to reference differ. At one extreme, there is the purely theoretical approach, basing the definition of optimality entirely on theoretical assumptions about speakers or hearers. This approach dominated the early years of REG research (Dale, 1989, 1992; Dale and Reiter, 1995; Reiter, 1990), as human data was not yet as widely available. At the other extreme, advances in Machine Learning and the availability of dedicated semantically transparent REG corpora, in particular the TUNA corpus (van Deemter et al., 2006), but also the GRE3D3 and GRE3D7 corpora (Viethen, 2011), have given rise to data-driven approaches where human data are used as input to learning algorithms, and system responses are automatically optimised to fit the data. For example, Theune et al. (2007) present, among others, a system which uses learnt cost functions for attributes, and always selects the distinguishing description with the lowest overall costs.

In between these two extremes, there is a wide range of more or less *empirically inspired* approaches where results from empirical studies are used to inform system design, and a formal definition of optimality is derived from a mixture of these results and theoretical considerations. Some approaches also use data-driven approaches to optimise certain aspects of otherwise *empirically inspired* systems. For example, Golland et al. (2010) propose a theoretically motivated system which incorporates learnt models of spatial relations into the REG process. To give another example, Kelleher and Kruijff (2006) present a variant of the IA which bases the proposed property order on empirical findings regarding the relative cognitive ease of using different types of properties.

The two dimensions of human-likeness vs. understandability and theoretically founded vs. data-driven interact in interesting ways. Optimising or evaluating an REG system using a semantically transparent corpus of human-generated REs is considerably cheaper than optimisation or evaluation based on understandability, as for the latter it is not possible to compile a corpus of evaluation data *a priori* – the REs an algorithm generates can only be

evaluated once they have, in fact, been generated (unless one wants to evaluate *a priori* all possible REs that may be generated in a given situation, which would be very costly). Thus, it is not surprising that data-driven approaches to REG currently show a strong tendency towards human-likeness as optimality criterion.

In this thesis, the central reference handling mechanism can be characterised as *empirically inspired*, as it is based on theoretical considerations founded on an extensive examination of existing empirical evidence on human production and interpretation of REs. Further, I will discuss how data-driven approaches tie into this general mechanism, particularly for modelling attributes (spatial relations, graded properties, and colour) This modular approach ensures transparency of the decision mechanism while making use of human data where available in order to increase the effectiveness of communication.

The distinction between theoretical, empirically inspired, and data-driven also holds for the evaluation of referential agents - while earlier systems were mainly evaluated using illustrative examples, based on formally defined optimality criteria, more recent systems are increasingly evaluated with respect to empirical findings, using both established empirical results and transparent REG corpora. As in data-driven system design, the evaluation according to human-likeness is inherently cheaper than that of understandability.

However, in a series of joint challenges on REG and NLG, such as the TUNA-REG challenge (Gatt et al., 2009, 2008) and the challenges for generating instructions in virtual environments (GIVE, Byron et al., 2009; Koller et al., 2010; Striegnitz et al., 2011) an organisational framework was established that allowed thorough evaluation of different REG systems according to both human-likeness and understandability.

Regarding evaluation, in this thesis I will evaluate generated expressions and RR in an example-based manner in order to demonstrate compliance with existing empirical results. Further, in two empirical studies, I will evaluate REs generated by the presented mechanism based on understandability using both task performance and human judgment measures, and the resolution of human-produced utterances by PRAGR. In addition, I will use experimentation on machine-machine interaction as it is used in AI research

(Spranger and Pauw, 2012; Steels et al., 2005) in order to further evaluate the potential of the PRAGR mechanism for achieving joint reference in REG and RR. Finally, further example-based evaluation will be performed to demonstrate the integration in a dialogue system for allowing referential grounding dialogues.

Based on the clarification of the perspective on optimality of REs taken in this thesis, I will now proceed to discuss the most central criterion of optimality, namely the concept of the distinguishing description, and how this concept relates to vagueness in reference. While the discussion mostly focuses on REG, it also has implications for RR, as is demonstrated for example by the work of Frank and Goodman (2014), which is described in more detail in Section 3.1.4.

3.1.3 Vagueness and the Distinguishing Description

The earliest and most fundamental criterion of optimality is derived directly from the definition of the classic REG paradigm of the first mention distinguishing description discussed in Section 2.1. If the scope of REG is restricted to distinguishing descriptions, then obviously the primary necessary criterion for judging the optimality of a description is whether it is, in fact, distinguishing. This criterion is so strongly enmeshed with classic REG that it is usually not questioned, and considered a necessary criterion any RE needs to fulfil, rather than part of the optimality definition. Moreover, this criterion is usually not defined empirically, but rather formally such that a description is considered to be distinguishing if it consists of a set of properties which all hold for the target object, but which do not all hold for any of the distractor objects (see also the more formal definition in Section 2.1).

This definition *per se* depends on the notion of crisp category membership associated with traditional REG approaches and causes difficulties when dealing with vagueness: if a speaker says *the large mouse* in a context where several mice are present, this may not be a distinguishing description in the strict sense, as the property LARGE holds for all mice to a certain degree. Nevertheless, the description may be sufficient for a listener to identify the

correct referent. Likewise, *the ball to the left of the box* may allow for the listener to easily identify which ball was meant, but if there is another ball to the left of the box, however marginal its position, that description does not fulfil the strict criterion that the set of selected properties may not fit any distractor object.

In order to determine how this mismatch between the binary nature of the distinguishing description criterion on the one hand, and vagueness effects on the other hand can be resolved, it is first necessary to gain a deeper understanding of the concept of vagueness in the context of cognition and communication. Therefore, in the following I will distinguish different perspectives towards the semantics of vagueness and clarify how the work presented in this thesis relates to these perspectives.

This will be followed by a discussion of different approaches to handling vagueness in reference, and to overcoming the limitations of the concept of the distinguishing description, and a positioning of this thesis with respect to the approaches discussed.

In Semantics, vagueness has traditionally been treated by incorporating borderline cases. For example, three-valued logic assumes that a category can have members, non-members, and borderline cases (van Rooij, 2011). This approach has been criticised as it still assumes crisp boundaries between, for example, members and borderline cases, while the defining characteristic of vagueness is the lack of crisp boundaries of any kind (van Rooij, 2011).

The approaches I am concerned with here are those which incorporate a more fine-grained notion of vagueness. van Rooij (2011) categorises these into degree-based approaches on the one hand, and delineation-based approaches on the other hand.

Degree-based Approaches

Degree-based approaches (e.g. Kennedy, 2007; Von Stechow, 1984; Zadeh, 1965) assume that each individual is a member of each category *to a certain, measurable, degree*, and thus can be assigned a graded membership value. For example, a given hue can be considered RED to a certain degree, and a

bright shade of red would have a high membership degree for the category RED and a value of 0 (or very close to 0) for YELLOW. A reddish hue with a strong yellow tinge, on the other hand, would have a lower membership degree for RED, and a membership degree for YELLOW which clearly deviates from 0. This kind of gradedness is also a case of vagueness, in the sense that there are objects for which, due to low membership grades in several categories, it is not *a priori* clear which category should be assigned to them.

Due to the graded category membership, sentences in which these categories are attributed to individuals, have truth conditions which can be stated in terms of degrees. This approach was formalised by the work of Zadeh (1965) who introduced the concept of *Fuzzy Sets* – sets which are not defined by member vs. non-member objects, but by membership functions which assign a degree of membership in the range [0,1] for each object, and logical operators which allow inferences over such fuzzy sets.

Degree based approaches have the advantage that they allow us to represent relevant cognitive phenomena of categorisation, as there is a vast body of empirical research which shows that humans perceive categories as having graded membership. Different individuals may be perceived as being more or less central to a given category.

These cognitive phenomena have been studied extensively following a major paradigm shift in the 1970ies which led away from previous assumptions that humans distinguish members from non-members of a category by means of essential features (*boundary*-based categorisation). Categories are now typically seen as characterised by an idealisation of what a perfect member would be, with membership depending on overall perceived similarity to this *prototype* (Mervis and Rosch, 1981).[2]

In experiments conducted by Posner and Keele (1968) and Reed (1972), participants faced with artificial stimuli with high variation in a category learning task tended to form idealised prototypes based on prior input which they then used for categorising new examples, thus giving rise to a graded

[2]Though see Frixione and Lieto (2012) for a discussion of representations via prototypes vs. stored exemplars.

membership structure.

While recent work on verbal and non-verbal direction categories has shown that boundaries do play a role in the representation and perception of at least some categories (Crawford et al., 2000; Huttenlocher et al., 1991; Klippel and Montello, 2007; Mast et al., 2014b), there is an overwhelming consensus that *most* conceptual and linguistic categories are to some degree based on prototypes. Research on both natural and artificial colour categories (Berlin and Kay, 1969; Rosch, 1973), shape (Rosch, 1973), projective terms (Gapp, 1995b; Zimmer et al., 1998), verbs (Coleman and Kay, 1981; Stamenković, 2011) has shown a variety of prototype effects for all these different domains.

These prototype effects can be separated into two different types: one regards vagueness, or ambiguous category membership, where due to the peripheral member-status in several categories, an individual can be ambiguous with respect to which category it belongs to (van Deemter, 2010, pp. 112–116). The other type of prototype effects regards typicality, where even for two clear and unambiguous members of a category, one can be considered a better representative of the category than the other (Rosch, 1973). For example, a bright red and a slightly more yellowish red may both be unambiguous members of the category RED, but the bright red may be considered more prototypical, i.e. more representative of the category as a whole. Thus, categories can exhibit gradedness without vagueness.

The classical fuzzy sets formalisation of graded category membership focuses on the first type, ambiguity of membership. However, prototype-based category models such as the one presented by Gapp (1995b) for projective terms represent both types, as gradedness within areas of non-overlap can be understood as a typicality effect, while gradedness in the areas of category overlap constitutes a vagueness effect.

Spranger and Pauw (2012) propose an extreme form of the degree-based view of categories such that, with increasing distance from the prototype, the acceptability rating for a category approaches, but never reaches 0 (see Section 3.1.4).

One problem of degree-based approaches to vagueness is that they do not

reflect very well the intersubjective aspects of language, i.e. that language is always the result of (implicit or explicit) negotiation within a community of speakers, and that the graded membership values of one speaker need not be the same as those of another speaker. The notion of a *measurable degree of category membership* gains an objectivist stance if one assumes it is possible to measure this degree of membership once and for all. On the other hand, if the degree of membership is taken to represent solely the perspective of a single speaker, it remains unclear how this relates to the linguistic community at large.

Delineation-based Approaches

As discussed in Section 2.3.3, humans adapt their use of language and underlying conceptualisations to each other in communication (Clark, 1996; Clark and Brennan, 1991; Garrod and Anderson, 1987; Wilkes-Gibbs and Clark, 1992). Thus whether or not to call an object *red* is subject to negotiation and change. This implies that a speaker can never be sure that the listener shares their conceptualisation of events, and that they make decisions about how to verbalise events in the face of this uncertainty.

Such issues are handled well by delineation based approaches (e.g. Kamp, 1975; Klein, 1980; Lassiter, 2011; Lawry and Tang, 2009; Lewis, 1970) which assume that while vague categories do in fact have clear-cut boundaries, these are usually unknown, or underspecified, thus constituting an aspect of uncertainty (Lawry and Tang, 2009). Though the idea of clear-cut boundaries has historically been associated with an objectivist stance, i.e. implying clear-cut boundaries in the real world, this approach can also be motivated from the point of view of adaptation to a (linguistic) community or a given situation.

To give an example, Lassiter (2011) assumes clear-cut, but unknown boundaries which are constrained by the norms of the linguistic community at large, and may be established by a group of interactants within a situation in the form of conceptual pacts within these constraints. Following these underlying assumptions, vague meaning is modelled in terms of probability

distributions over possible languages. As each possible language determines clear boundaries for each category, the meaning of a category is ultimately modelled in terms of a probability distribution over the category boundaries. Within a conversation, interactants may engage in a grounding process where this probability distribution is successively narrowed down until a more or less crisp category for the given interaction emerges.

Based on this approach, McMahan and Stone (2015) present a mechanism for learning colour concepts under the assumption of situationally dependent crisp boundaries with uncertain positions. Meo et al. (2014) apply this approach in a system for context-dependent generation and resolution of colour REs.

This demonstrates how non-objectivist delineation-based approaches assume clear-cut boundaries which are highly context-dependent and often underdetermined. For example, in every specific situation, there is a unique cut-off point between TALL and SHORT, but its precise identity may be unknown to the interactants, and is subject to negotiation, hence it initially only exists in terms of constraints or probability distribution over possible boundaries. In contrast to degree-based approaches which focus on the individual perception and explain cognitive effects of gradedness, delineation-based approaches are better suited to explain the intersubjective aspects of semantics.

Positioning of this Thesis

Intuitively, it is plausible to assume that delineation-based approaches in the *Semantics* sense correspond to boundary-based representations in the *Cognitive Science* sense, while degree-based approaches correspond to prototype-based representations. However, this is not necessarily true. There are a number of approaches combining prototypes with delineation-based approaches by deriving category boundaries from prototypes. For example, Gärdenfors (2004b) suggests deriving crisp category boundaries from prototypes by *Voronoi tessellation*, as will be discussed in detail in Chapter 4. Douven et al. (2013) extend this approach to yield a three-valued logic (see

Chapter 4).

Eyre and Lawry (2014) propose to combine prototypes with probabilistic boundaries similar to the possible languages approach by Lassiter (2011). In their work, probabilistic boundaries are derived from prototype representations based on the distance of an individual to the prototype, thus combining prototype representations with delineation-based probabilistic Semantics. This approach has the advantage of capturing both the graded nature of categories in perception, and the intersubjective nature of categories in communication which are subject to differences and negotiation. Thus, the approach suggested by Eyre and Lawry (2014) functions as a bridge between the cognitive and the communicative perspectives of categorisation, binding them together via the intuitive assumption that the more acceptable one member of the linguistic community considers a category to be for a given individual, the more likely it is that another member of the community will share the same conceptualisation, and will thus accept the category as an adequate description of the individual. This is the perspective that I take in this work, as will be explained in more detail in Section 3.2.

Notes on Vagueness and the Distinguishing Description

With respect to the concept of the distinguishing description, the two perspectives on vagueness described above have slightly different implications. The degree-based perspective implies that when vagueness is involved, there literally is no such thing as a distinguishing description, but each description can only be more or less distinguishing. Thus, going back to the example of the room full of mice, if a speaker says *the large mouse* in a context where several mice are present, this description is by definition not distinguishing in an absolute sense, as the property LARGE holds for all mice to a certain degree. However, it may have a higher or lower degree of discriminatory power (see below). If all other mice are very small, and thus members of the category LARGE only to a very low degree, the description would be considered *highly discriminatory*. If, on the other hand, there are other mice present which are members of the category LARGE to an almost equally high

degree as the target mouse that is being described, the description would be considered to be only *weakly discriminatory*.

The delineation-based perspective, on the other hand, assumes that there is a clear distinction as to whether any given description is distinguishing or not. However, given the incomplete information of a speaker, the best they can do is to *estimate* the chances, or probability, that this is in fact the case. Thus, in the face of one extremely large mouse in a room full of fairly small mice, they might estimate that probability to be very high, reaching almost 1 perhaps, while in the presence of several fairly large mice, the probability of the description *the large mouse* being distinguishing will be estimated to be much lower.

Finally, an aspect of uncertainty which is not considered in Semantics, but is also highly relevant to human machine interaction, is uncertainty that can arise from unreliable sensory data. For example, even if an agent had a clear definition of the boundaries of the category RED, the available values of hue, saturation and lightness for a certain object may vary greatly due to lighting conditions in the environment, posing problems for categorisation. While this is certainly a highly relevant aspect of uncertainty, especially in the face of human-robot interaction, it will not be treated explicitly in this thesis. I do however consider the overall approach of PRAGR to be fairly robust with respect to such perceptual uncertainty, an issue which will be addressed in Chapter 6.

3.1.4 Handling Vagueness in Reference

Now that I have clarified the different perspectives on vagueness, I will continue to discuss different ways to incorporate vagueness into the generation and/or resolution of REs. Three underlying approaches for handling vagueness and uncertainty in reference with *grounded symbols* (Steels, 2008) can be distinguished which I will term (I) the Independent Decision Approach (IDA), (II) the Global Decision Approach (GDA), and (III) the Modular Decision Approach (MDA).

The Independent Decision Approach

The IDA is the most straightforward adaptation of classic REG to vagueness. While it considers issues of vagueness and uncertainty in decisions regarding individual properties, it makes independent incremental decisions for each property. This approach is pursued by van Deemter (2006), Horacek (2005), and Kelleher and Kruijff (2005) for generating REs, and by Gorniak and Roy (2004) for resolving REs. In the IDA, for each property a separate, crisp decision is made (whether to add the property in REG, and how to prune the distractor set in RR) as in the classic approach to reference, but instead of performing a priori, crisp property assignment, properties are assigned flexibly based on quantitative information representing vagueness and/or uncertainty.

Van Deemter (2006) proposes an approach which transforms numeric values into sets of inequalities, e.g. if a mouse is 6 cm long, this may be transformed to $\{> 4$ cm, < 8 cm, < 10 cm$\}$. The algorithm then generates superlative expressions for sets of objects, such as *the largest 3 mice*, by checking whether there is an inequality which all target objects share, which is not shared by any distractors. Thus, the same mouse may be called *large* or *small*, depending on the physical and linguistic context: which other mice are present, which other mice are part of the target set, and which other properties have already been selected.

Horacek (2005) considers the probability of correct interpretation of each given property based on term knowledge, perceptual risks, and how likely the listener is to follow the conceptualisation proposed by using the property for the given object. A property is only added to the overall RE if it exceeds a threshold probability of correct interpretation.

Gorniak and Roy (2004) address issues of interpreting combinatory spatial REs such as *the lowest purple on the right hand side* by using incremental filters. Following a predetermined order, at each step a set of potential referents of each property is filtered out, based on perceptual information on the objects.

Kelleher and Kruijff (2006) apply the IDA to REG with spatial relations. They divide the acceptability of spatial relations for distractors into three distinct categories: (a) better fit than the target, in which case the relation would not discriminate the target from the distractor at all, (b) acceptable, but worse fit than the target, in which case the relation discriminates the target from the distractor *relatively*, and (c) not acceptable, in which case the relation discriminates the target from the distractor *absolutely*. Based on this distinction, they assume a cognitively motivated preference for absolute discrimination over relative discrimination, thus relying only on relatively discriminating relations if no absolutely discriminating relations can be found.

The Global Decision Approach

The GDA obtains a single, integrated model of attribute selection in REG by application of Machine Learning to a semantically transparent corpus of REs paired with perceptual scene representations. Thus, instead of making a crisp decision for each property, a holistic decision for an entire description is made based on low-level perceptual features.

Tellex et al. (2011, 2014) train a probabilistic graphical model on a semantically transparent corpus. The model aims to maximise the probability of a set of groundings (mappings of constituents to real-world objects, locations, or paths), given a parsed natural language command and a perceptual model of the environment. The model incorporates factorisation of probabilities based on syntactic structure. Word meanings are modelled using a total of $147,274$ binary features representing predetermined labels (e.g., *truck*) and a range of simple geometric relations (e.g., distance). The search space for identifying potential groundings is restricted using topographical maps to identify salient candidates.

Engonopoulos and Koller (2014) use a log linear model for evaluating the discriminatory power of an RE for an object x based on a score for its fit as a weighted sum of features. Here, features are not necessarily models which determine acceptability for a given concept. They may also be features

which cover only a certain aspect of one or more concepts, e.g. a feature might consist of a distance score which may influence the acceptability of any projective relation, and relations such as NEAR or AWAY FROM, thus differentiating this approach from the MDA described below.

Golland et al. (2010) present a probabilistic, game-theoretic model for the selection of reference objects and spatial relations in REG. A *rational speaker* estimates the expected utility of an RE in terms of the probability that a listener would select the correct target for this RE. The authors suggest a learnt listener model, the parameters of which are learnt based on a semantically transparent corpus of human produced utterances by maximising the likelihood of the targets, given the spatial relations and reference objects chosen by the humans. Thus, in a sense, the model presented by Golland et al. (2010) learns a measure of understandability based on human produced data. While complex descriptions consisting of chained relational descriptions are addressed, the model remains restricted to spatial relations and is not expanded to cover a wider range of properties.

The Modular Decision Approach

The MDA provides some combinatory means of determining the suitability of an entire RE consisting of multiple properties, based on an evaluation function of individual properties. Unlike the GDA, which attempts to learn the parameters for the entire decision process in an integrated fashion, the MDA models acceptability of features independently of the referential situation. For REG and RR, a combinatory scoring mechanism is used which takes into account property acceptability, situational context, and referential goals. The MDA has become popular in recent work, and a number of systems following this approach have been proposed during the time taken to complete this thesis, the most relevant of which I will briefly introduce here.

Frank and Goodman (2012) present a Bayesian Model of pragmatic reasoning in language games that demonstrates how probabilistic inferences may improve RR. They model the probability of a speaker using a particular description to refer to a given object as the *surprisal* of this description – i.e., its

both.

Meo et al. (2014) present a method for context-dependent generation and resolution of colour REs using crisp boundaries with uncertain positions. Their model learns the probability distribution of category boundaries from descriptions of colour swatches elicited from human subjects. Based on this distribution, it aims to jointly maximise Discriminatory Power and Acceptability of colour descriptions in the given context by estimating the likelihood of category boundaries which include the target object, but exclude the distractor. Meo et al. (2014) integrate diverging preferences for different colour terms by evaluating the relative frequency of a category, given that it is true of a point in colour space. The work by Meo et al. (2014) however does not deal with the generation or interpretation of complex REs with different property types.

Spranger and Pauw (2012) propose a system for REG and RR which uses graded acceptability values for properties based on similarity to prototypes. With increasing distance from the prototype, the acceptability rating for a category approaches, but never reaches 0. The mechanism is capable of dealing with cases of non pareto-optimal property combinations by using multiplicative scoring for combined properties. As Spranger and Pauw (2012) demonstrate in language game experiments with robots, their *lenient* approach to vague semantics allows the agents to overcome perceptual deviation, and thus yields higher communicative success than a crisp approach based on Voronoi tessellation.

Funakoshi et al. (2012) employ Bayesian networks for RR which estimate the probability with which a given RE refers to a particular entity. The probability that a given word or multi-word utterance refers to a particular object is dependent, among other factors, on the assumption of a reference domain d. A reference domain may be the entire scene, or a group of objects which stands out perceptually or was referred to earlier in the dialogue. Thus, an expression such as *the right one* can refer to the rightmost object in the entire scene, or to the rightmost of a certain subset of objects, depending on what the relevant reference domain is, allowing the understanding of context dependent properties. The authors note that the application of this or similar

approaches to REG would be desirable (Funakoshi et al., 2012, 244), but no specific suggestions beyond full search are made for this.

Roy (2002) proposes the DESCRIBER system which is capable of learning perceptually grounded categories and using them for REG. The system learns meaning of linguistic terms from pairings of images with a target object and linguistic descriptions of that object. Words are bound to perceptual features based on clustering, and Acceptability is modelled as a multivariate Gaussian function based on the selected features.

REG in this system is driven by, and integrated with, surface realisation. Based on learnt syntactic constraints over the formed word classes, for a given description length the most likely sequence of word classes is estimated. Then for each word class, the most likely word is determined by comparing the fit of the utterance for the target object to its fit for possible distractor objects. The fit of the whole description is the product of the fit of each word in the description. The best description is the one which jointly maximises syntactic and contextual constraints.

Comparison of Approaches

The IDA clearly has the advantage that it can rely on the concept of distinguishing description, and the existing REG algorithms. Thus, it can provide a simple, computationally efficient method for handling symbol grounding and vagueness in reference. On the other hand, this also leads to some problems. Due to the independent nature of decisions, for each property, some minimal discrimination threshold has to be set – the granularity of the inequalities (van Deemter, 2006), the minimal discrimination improvement (Horacek, 2005), or the filtering threshold of the composer functions (Gorniak and Roy, 2004). Beyond this, the size of the contrast in a given dimension is not considered. Thus, in situations like Figure 2.1a (p. 22) above, highly contrastive properties would be handled exactly like those just above the threshold.

This solution can cause problems for handling combinatory descriptions. As van Deemter (2006) notes, handling combinations such as *the tall fat*

giraffe which are not pareto-optimal – i.e., the giraffe in question is neither the tallest, nor the fattest, but the best candidate for the combination of both – causes difficulties for this approach, as quantitative information regarding the size of contrast in each domain is lost at the combination stage. In the case of reference resolution (Gorniak and Roy, 2004), this can lead to faulty intermediate pruning decisions and thus require search to identify the most plausible interpretation. For example, despite being able to handle spatial relations like BELOW and TO THE RIGHT OF, the model proposed by Gorniak and Roy cannot handle descriptions such as *below and to the right* (Gorniak and Roy, 2004, p.442). In contrast, the MDA and GDA handle such *tall fat giraffe* cases straightforwardly and without backtracking, both in REG and RR. Due to the inherent inability of the IDA to handle combinatory descriptions in a satisfying manner, I conclude that a binary notion of discrimination – either a description discriminates the target from a given (set of) distractor(s), or it does not – is not sufficient for handling reference in the face of vagueness and perceptual deviation, as they occur in realistic settings.

The GDA has the advantage that the integrated suitability function takes into account any dependencies between different property models, and can be trained directly from description data. However, the latter can also be seen as a disadvantage, as this requires semantically transparent training data that distributes across all objects and all potential contexts, in order to learn the integrated suitability function. Thus, for each new application scenario, a new training corpus of considerable size is needed which is problematic due to the high cost involved in gathering semantically transparent corpora.

The MDA, on the other hand, provides mutually separated acceptability functions for individual properties. Models can be manually designed to handle the specifics of each feature domain while machine learning can be applied to optimise parameters of the individual components. Likewise, models of linguistic and visual salience can be manually designed and optimised with machine learning (Kelleher, 2011). Each model can be trained separately, based on more easily available acceptability judgments on individual attributes. It is to be expected that most models can be transferred

to new application scenarios, as the integrated scoring mechanism handles the integration of scores under consideration of the specific context. It may be necessary to learn weights for integrating features for new application scenarios, but a relatively small corpus should be sufficient for this purpose.

For these reasons, with PRAGR I follow the MDA, providing individual property models and a probabilistic integrated mechanism to calculate appropriateness of descriptions based on the acceptability of individual properties. Spranger and Pauw (2012) provide successful evaluation of robot-robot communication with spatial relations using this approach, while Meo et al. (2014) demonstrate the ability of their approach to mimic human decisions for discriminatory colour naming. In an example-based evaluation, Kelleher (2011) demonstrates the ability of his system to successfully handle one-anaphora. While all of these authors in principle enable combinatory evaluation of complex REs with multiple property domains, I am not aware of any evaluation which empirically demonstrates their ability to successfully generate and/or interpret such complex REs with several different vague property domains and spatial relations. The aim of this thesis is to fill this gap by developing and evaluating such an integrated reference handling mechanism.

Discriminatory Power in the Context of Vagueness and Probability

Any REG system following the MDA requires some global measure of goodness of an RE. As I have discussed above, the concept of the distinguishing description does not suffice, as the question whether a given property distinguishes a target from a given distractor is not answered in a binary way by the MDA. A key concept that has been used in different variants to this end is the concept of *Discriminatory Power*. Discriminatory Power is a measure of how much a description contributes to distinguishing the target from the distractor set.

Within the paradigm of crisp categories, the concept of Discriminatory Power has been used by Dale (1989, 1992) as the foundation for the Greedy Heuristic Algorithm (GH). Dale (1989) defines the Discriminatory Power $F \in [0, 1]$ of an attribute-value pair $\langle a, v \rangle$ which is true of the intended

referent.F can be determined by the equation:

$$F(\langle a, v\rangle) = \frac{N-n}{N-1} \tag{3.1}$$

where N is the number of objects in the distractor set, and n the number of objects in the distractor set for which $\langle a, v\rangle$ is true. A value of $F = 0$ means the attribute-value pair contributes nothing to discriminating the target, while with increasing F the contribution of the property is larger, and a value of $F = 1$ means it uniquely identifies the target. Based on this definition, the GH iteratively selects the attribute-value pair with the highest discriminatory power based on the set of remaining distractors, until a discriminating description has been found.

A probabilistic variant of Discriminatory Power has come up in the game-theoretic approach to reference which applies the idea of Wittgensteinian language games (Wittgenstein, 1953) to reference [3]. Game Theory is a framework used within a broad range of disciplines, such as biology, psychology, economics and computer science. It encompasses "the study of mathematical models of conflict and cooperation between intelligent rational decision-makers" (Myerson, 1991). A core idea of Game Theory is hypothetical reasoning and evaluation of the opponent's or collaborator's potential actions by an agent in order to inform their own decision making. The domain of reference is well suited to this framework, as it involves the necessity of two individuals to make decisions (generating and resolving REs) within a co-operative setting with a joint goal. Moreover, as discussed above, humans have been shown to engage in this kind of reasoning about their interlocutor to some degree (e.g., Frank and Goodman, 2014).

Researchers following the MDA for integrating vagueness and uncertainty into REG have used different variants of Discriminatory Power. In the work by Spranger and Pauw (2012), Discriminatory Power is the distance between the acceptability of a combined score for the target, and for the closest dis-

[3]The work by Spranger and Pauw (2012), which also follows the language game approach, will be discussed below

tractor. For Meo et al. (2014). Discriminatory Power is the probability that category boundaries fall in such a way that the description includes the referent, but excludes the distractor.

In this thesis, I will build on these ideas and provide a definition of Discriminatory Power which includes vagueness. I will provide an explicit probabilistic model which creates a link between graded acceptability of categories and the discriminatory power and the appropriateness of a description. Further, following the fundamentals of Bayesian statistics, I will provide an account of the relationship between the acceptability and discriminatory power of simple descriptions to those of complex descriptions involving several objects and properties. This approach will be spelt out in detail in Section 3.2, though the underlying idea is as follows: if a given description suits the target object very well (*the red ball* for an object which is a prototypical ball coloured a prototypical shade of red), but does not suit any distractor well at all (e.g. all distractors are green boxes) the Discriminatory Power of the description is high. The less well the description suits the target object (e.g. the object is a less typical ball, or has a non-prototypical shade of red), and the more distractor objects it suits better (e.g. there are one or more objects in the scene which are also red balls to some degree), the lower the Discriminatory Power of the expression. While the probabilistic mechanism presented here does not inherently rely on a particular way of representing uncertainty, in this thesis I will use a prototype-based approach which will be described in Chapter 4.

3.1.5 Further Aspects of Optimality

Whether crisp or not, the principle of the distinguishing description or Discriminatory Power on its own is not sufficient to determine the optimal set of properties. Theoretically, mentioning all properties that are acceptable for a given referent (to a certain degree) would always lead to a distinguishing description, unless there was a distractor present for which all these properties are true as well – in which case it would be impossible to create a distinguishing description (Krahmer and van Deemter, 2012, p. 178).

Therefore, all approaches to REG incorporate further criteria of optimality. As discussed above, the criterion of human-likeness is not considered here. A further aspect mentioned above is salience, which does not constitute an optimality criterion in the direct sense – it does not make sense to require REs to be more or less salient, in fact it is not even clear what that would mean. However, salience is relevant in that, similarly to vagueness, it constitutes a concept that should be taken into account in the optimality definition. In other words, in a definition of optimality that is suitable to situated interaction, the salience of different objects mentioned in a descriptions should impact the evaluation of that description. Thus, if a description mentions a highly salient object as a reference object, it should be considered better than a description that mentions an object with low salience. For the sake of the current argument, it suffices to note that the optimality definition presented in this work provides a means to incorporate object salience. The implications of this will be discussed in detail in Section 5.1.4.

Finally, there is the criterion of brevity. Based on the ideal of understandability which is adopted in this thesis, early REG work such as the Full Brevity Algorithm (FB) (Dale, 1989), aimed at reducing the cognitive load of the listener. Based on Grice's Conversational Maxim of Quantity: "do not make your contribution more informative than is required" (Grice, 1975, p. 45), this approach attempts to find the distinguishing description with the least number of properties, i.e. the shortest description[4]. The underlying assumptions are that longer descriptions are harder to process *per se*, and that by violating the Maxim of Quantity, they may give rise to unintended conversational implicatures, thereby causing confusion. While the FB performs a full search, always yielding the shortest distinguishing description at the price of high computational complexity, the GH approximates this by successively adding the most informative property to the description, thus exploiting the concept of Discriminatory Power discussed above.

[4]Obviously, this is a simplified view of brevity of descriptions, as the expression of some properties may be more verbose than that of others. However, since surface realisation is not the focus of this work, I will adopt this view of brevity for the sake of the argument.

The use of brevity as an optimality criterion has received much criticism, based on empirical evidence that humans in fact tend to produce redundant REs with more properties than are strictly needed for identifying the target object (Mangold and Pobel, 1988; Olson and Ford, 1975; Pechmann, 1984, 1989). Pobel et al. (1988) give an overview of empirical research which shows that speakers over-specify descriptions either in order to make production easier for themselves, or to make reference resolution easier for the listener. With respect to understandability, further empirical work shows that under certain conditions, over-specification can speed up identification (Arts et al., 2011; Engelhardt et al., 2006; Sonnenschein, 1984). Specifically, over-specification seems to be helpful if the additional property adds information which reduces the search space, or helps the listener complete a mental image of the referent (Arts et al., 2011).

To conclude, while brevity is obviously a relevant aspect of optimality in REG, it is by no means the single most important aspect. In this work, I will use brevity as a secondary criterion which may be used to enforce a decision between otherwise equally acceptable expressions. Moreover, as will be shown in the following section, by virtue of probabilistically representing category acceptability, in PRAGR any property which is less than perfect for a given object automatically increases the cost (or lowers the appropriateness) of the RE as a whole, thus increasing the likelihood of the system choosing short descriptions, if these provide sufficient discriminatory power.

3.2 Probabilistic Mechanism

In the last sections of this chapter, I have provided motivation for defining the optimality of an RE in terms of understandability, by using a graded measure of Discriminatory Power and allowing for the integration of salience effects into the core mechanism. In the following, I will present the core mechanism of PRAGR, before turning to more detailed descriptions of the identified challenges for REG in the next chapters.

3.2.1 Core Concepts of PRAGR

The core concepts of PRAGR are Acceptability and Discriminatory Power which I will define in the following.

The meaning of **Acceptability** is twofold. On the one hand, the Acceptability of a description reflects the graded acceptability of the description for the given object, as perceived by the agent, i.e. the willingness of the agent themself to accept RED as a valid conceptualisation of the object in question. This corresponds to the degree-based notion of vagueness. On the other hand, it can be viewed as the conditional probability $P(D|x)$ that the interlocutor would accept description D for object x, following the delineation-based approach. Thus, the link between both approaches is provided by the intuitive notion that the more central the perceived feature of the object is in the individual's model of the qualitative property, the higher the chance that other individuals would accept this qualitative property for the object in question. An object which an agent perceives as a perfectly prototypical ball in a perfectly prototypical shade of red is more likely to allow another agent to follow the conceptualisation as {RED(x), BALL(x)} than an object which only vaguely looks like a ball to this agent, and appears to have a marginal shade of red.

The reason for this two-fold interpretation is that on the one hand, it pertains to the conceptual system of an individual where categories can have graded membership, and on the other hand it pertains to hypotheses an individual will make regarding communicative options and their likelihood for success. Thus, a speaker may consider a colour a good member of the category MAUVE, but still estimate chances of their interlocutor following this conceptualisation as low, because they know the interlocutor to be colour blind or not very knowledgeable regarding sophisticated colour terms. While in the current implementation, a direct transfer from perceived graded category membership to evaluation of communicative potential is assumed, in principle a further step of processing needs to be assumed which allows for divergence between those two aspects.

As far as the reference handling mechanism itself is concerned, the reas-

ons for reduced acceptability are irrelevant. Acceptability thus provides the interface between the individual property models which will be described in more detail in Chapter 4, and the reference handling component. From a technical perspective, Acceptability provides the interface between knowledge representation and content selection. On the theoretical side, Acceptability provides the link between individual cognition and categorisation on the one hand, and social interaction and communication on the other hand. Using Acceptability as an interface between property models and reference handling mechanism ensures that the system remains modular, allowing different models of features depending on domain characteristics. For example, there is good reason to assume that graded properties need to be modelled differently from colour. Also, this modularity enables compatibility with different applied systems which may use different property models depending on available resources, e.g. different kinds of sensors or perceptual-level processing. At the same time, by using a probability value rather than a binary distinction, the core reference handler retains much of the relevant information needed in order to make strategic use of graded category acceptability for REG.

For the sake of spelling out the core mechanism of PRAGR and the probabilistic optimality definition for REG, we can simply assume at this point that $P(D|x) \in [0, 1]$ is defined for each primitive feature, i.e., for each object and feature the mechanism is provided with a number between 0 and 1 that determines the respective feature acceptability for this object.

Discriminatory Power, on the other hand, measures the degree to which a description D can discriminate the intended target object x from its distractors, in terms of the degree-based approach, or the conditional probability $P(x|D)$ that the listener would identify the correct object x, given description D, from a delineation-based perspective. Using the well-known Bayes' theorem, it is possible to express $P(x|D)$ in terms of $P(D|x)$, therefore making $P(D|x)$ the main factor of the model:

$$P(x|D) = \frac{P(D|x)P(x)}{P(D)}, \tag{3.2}$$

where $P(x)$ is the prior probability of the object, i.e. the probability that, if one randomly chose an object of the given scene, one would pick x. In a simple model, we can assume an equal probability of being selected for each object, thus defaulting to $P(x) = \frac{1}{N}$, where N equals the number of objects in the context. In the extension of the model presented in Section 5.1.4, $P(x)$ incorporates the *Salience* of the object, as more salient objects will be noticed more easily and therefore have a higher chance of being randomly selected.

The prior probability of the description, $P(D)$, gives the probability that the description D suits a randomly chosen object from the context C:

$$P(D) = \sum_{x_i \in C} P(D|x_i)P(x_i), \tag{3.3}$$

which, under the simplifying assumption of equal probability of all objects (i.e. ignoring salience) amounts to

$$P(D) = \frac{\sum_{x_i \in C} P(D|x_i)}{N}. \tag{3.4}$$

To summarise the definition of $P(x|D)$, the better D fits x, and the less well D fits the other objects in the context, the higher the Discriminatory Power $P(x|D)$ of the description for this object.

3.2.2 Referring Expression Generation

Based on this model and considering the thoughts on Discriminatory Power discussed above, one may assume that the speaker intends to optimise Discriminatory Power, thus choosing the most discriminating description. However, this approach would have the undesired side effect of aiming to determine the one description that cannot be interpreted wrongly. Therefore, all possible features of an object would be exploited to compile a description, even if they were not very good descriptions for the object *per se*, yielding a long and inadequate description of the object. Human speakers prefer descriptions which suit the target object well, as shown for example in the

preference of subjects for prototypical spatial relations (Carlson and Hill, 2009). Further, as the goal of producing an RE is not only discrimination, but also the achievement of grounding, i.e., an agreement on how to conceptualise the objects in the environment, this goal needs to be taken into consideration, too.

Therefore, we need to counterweight $P(x|D)$ by Acceptability, assuming that a speaker intends to produce a description that is both discriminating and acceptable. This leads to the concept of Appropriateness as a weighted average of Acceptability and Discriminatory Power, yielding the optimality definition:

$$D^* := \arg\max \Big((1-\alpha) P(x|D) + \alpha P(D|x) \Big) \tag{3.5}$$

Thus, the best description D^* is defined as that description which maximises the weighted combination of Acceptability $P(D|x)$ and Discriminatory Power $P(x|D)$, where the model parameter α determines the relative weight of Acceptability vs. Discriminatory Power. A higher value for α corresponds to a stronger weight of acceptability.

3.2.3 Reference Resolution

Regarding RR, in this work I assume a naive listener who, given a description D, selects as best referent x^* the object for which D is most acceptable:

$$x^* := \arg\max_x P(D|x) \tag{3.6}$$

Frank and Goodman (2014) show that a more sophisticated listener may improve communicative success by engaging in counterfactual reasoning. As the focus of this work lies on REG, this possibility is not further explored here.

3.2.4 Complex Descriptions

A key advantage of using a probabilistic approach is the straightforward extension to complex descriptions. In the probabilistic model presented in

this thesis, descriptions are sets of tuples that relate feature domains (COLOR, SIZE, LOCATION, etc., called *features* henceforth) to respective feature values (RED, LARGE, LEFT) for an object in the scene. Technically speaking, a description is a set of triples

$$\mathcal{D} := \{(o_1, f_{1,1}, v_{1,1}), (o_1, f_{1,2}, v_{1,2}), \ldots (o_n, f_{n,m}, v_{n,m})\}$$

that relates scene objects $o_1, \ldots o_n$, feature domains $f_{i,j}$ and feature values $v_{i,j}$.

The probabilistic model can be applied to descriptions of arbitrary complexity, i.e., any size of set of object/feature/value mappings. A prerequisite for this is to assume that the acceptability of different features is stochastically independent. For example, the probability of a human accepting that a door is RED (COLOUR feature) is assumed to be independent of the probability of accepting that the door is LARGE (SIZE feature). Obviously, this assumption of independence is an idealisation, as for example the feature SIZE may well correlate to some degree with other features such as HEIGHT for certain types of objects. However, for the sake of practical implementation the assumption of independence is a reasonable choice to make. It allows complex descriptions to be separated by feature domains, while the practical consequences of potential feature correlations for the task of REG are of limited relevance for the work presented in this thesis. Based on this assumption of independence, $P(D|x)$ of arbitrarily complex descriptions can be determined by multiplying single object/feature domain/value triples: $P(D|x) = P(f_0|x) \cdot \ldots \cdot P(f_n|x)$, where each f_i stands for a single object/feature domain/value triple.

When looking at spatial descriptions, it is also necessary to consider descriptions involving multiple objects, as the reference object needs to be described, too. As an example, if the speaker says *the ball to the right of the red box*, the description of the reference object – *the red box* – becomes part of the whole description. As this extension poses a number of interesting challenges, it will be discussed in detail in Chapter 5.

3.2.5 Some Notes on Acceptability

In the light of recent research, some design decisions made in the development of PRAGR regarding the handling of Acceptability need to be justified. In particular, the fact that Acceptability is an amalgamation of two distinct probabilistic concepts. Secondly, regarding the way in which Acceptability values of 0 are handled in this thesis.

Dual Function of Acceptability

In this work, Acceptability to some degree is an amalgamation of two distinct probabilistic concepts. Strictly speaking, one should distinguish $P(D_{said}|x)$ and $P(D_{true}|x)$, where $P(D_{said}|x)$ is the value to be used in the equation for deriving Discriminatory Power given above, while being itself derived from $P(D_{true}|x)$.

In line with this perspective, Funakoshi et al. (2012) do not use the results from the property models directly for deriving Discriminatory Power, but derive a probability distribution for $P(d_{said}|x)$ by normalising over all possible properties for an object, plus a value for Ω indicating the use of any other feature. This is technically a cleaner solution than using non-normalised probabilities as done with PRAGR.

However, this normalisation leads to some counter-intuitive effects that impact the usefulness of generated descriptions: $P(d_{said}|x)$ as derived from $P(d_{true}|x)$ by normalisation over all possible properties of an objects yields lower values if most available properties for the object have high Acceptability values. This means that two objects with identical un-normalised Acceptability may receive vastly different values for $P(d_{said}|x)$, a fact which may lead to over- or underestimating Discriminatory Power of a given description: assuming the unlikely scenario of a scene of exactly two objects, where one has acc(RED) = 0.9 and acc(LARGE) = 0.9, but another has acc(RED)= 0.1 and acc(LARGE)= 0.1, with no other properties specified, both have $P(\text{RED}|x) = 0.5$ and $P(\text{LARGE}|x) = 0.5$, thus *the large red object* would be deemed not discriminatory at all in this scenario.

In particular when generating REs with relations, this may lead to un-

derestimating the potential for discriminating a reference object which has a large number of highly suitable features and selecting instead a reference object with less suitable features.

The problem can be ameliorated by ascertaining for each attribute that the Acceptability values of all its possible values always add up to the same score (e.g. for the domain SIZE the Acceptability values of LARGE and SMALL always add up to 1). In terms of Conceptual Spaces this implies that prototypes of categories are evenly distributed within a conceptual space. In practice, this requirement is hard to guarantee, e.g., for the COLOUR domain. Further, if this is achieved, the normalisation has no relevant effect. Finally, from a semantic perspective, the normalisation performed by Funakoshi et al. (2012) implies that only one property would be used to describe an object which does not seem reasonable.

For these reasons, I have chosen to follow the assumption that $P(D_{said}|x)$ correlates linearly with $P(D_{true}|x)$, implying that objects which have no good properties will be less likely to be described at all. Technically in order to derive $P(D_{said}|x)$, $P(D_{true}|x)$ should be normalised by an unknown constant C which can be omitted due to having no relevant effect.

Throughout this thesis, I use the term Acceptability, omitting $P(D_{said}|x)$ for convenience, although the distinction may need to be made explicit for future versions of PRAGR, for example if different preferences for individual attributes or properties are to be integrated, as has been done in similar work by Meo et al. (2014).

Handling Acceptability Values of 0

As discussed above,Acceptability values are combined multiplicatively, thus $P(D|x) = P(d_1, \ldots, d_n|x) = P(d_1|x) \cdot \ldots \cdot P(d_n|x)$. This has the effect that containing one concept with an Acceptability of 0 sets the acceptability of the entire description to 0. While in principle reasonable – after all, calling a ball which can under no circumstances whatsoever be considered to be RED *the red ball* is blatantly unacceptable. However, this approach obscures the fact that the description *the red ball* is still more likely to fit a green ball

than it is to fit a green poodle, a distinction which is important for robust RR.

Engonopoulos and Koller (2014) combine Acceptability scores by weighted addition, where the weighting reflects how sensitive a particular feature is to errors – an error on type may be more drastic than one on size. In RR, the additive approach ensures that a single property of zero acceptability does not prevent the system from determining the most likely referent. For REG, the additive combination implies that the presence of very badly fitting properties has a fairly small influence over the description as a whole. In the extreme case, this leads to a counter-intuitive effect where, given any description D, adding a property p which has a very low acceptability (e.g. e^{-8}) will increase the acceptability of that description. Regarding Discriminatory Power, adding a property with Acceptability 0 for the target and for all distractors (i.e. a property which doesn't fit any object in the scene), would not impact discriminatory power negatively, as the acceptability of the combined description for all objects would remain exactly the same. In practice, this is not a big problem, as properties with Acceptability of 0 can be excluded from consideration *a priori*, but conceptually it is not an ideal solution.

Funakoshi et al. (2012) ameliorate the problem of 0 values for Acceptability by adding an intermediary step between concept and word, where each word is given a small probability ϵ of having been caused by some unknown OTHER concept with Ω being the likelihood that the object is being described using this unknown concept. Thus, a small value $\epsilon \cdot P(\Omega|x)$ is added to each individual word acceptability before performing the multiplication. This ensures that the acceptability of a complex description is never 0, thus still allowing differentiation of an object for which one word of the description does not fit at all from one for which several words do not fit at all. A similar solution can be achieved by setting a minimum value for acceptability and assigning all concepts that are not acceptable this minimal value. This approach has been used for the evaluation studies presented in Chapter 6.

3.3 Evaluating the Basic PRAGR Mechanism

Now that the basic PRAGR mechanism has been introduced, I will proceed to perform an example-based evaluation of PRAGR in a number of simple scenarios in order to demonstrate the ability of PRAGR to generate useful REs in the face of vagueness.

In order to generate REs, PRAGR requires a probabilistic mapping of objects to properties – the acceptability of the properties for the given object. While the quality of the generated REs is impacted by the quality of the property models, PRAGR itself does not impose restrictions on the way this mapping is achieved, as long as each property-object pair is mapped onto a number in the range [0,1] which represents the probability of the interactant accepting the given property for the object.

Therefore, for the purpose of evaluating the basic mechanism with illustrative examples, it is sufficient to describe two simple property domains. An extensive discussion of the approach towards property modelling used in this thesis, and an overview of all models used for further evaluation will be given in Chapter 4.

As discussed in Section 3.1.4, PRAGR follows the Modular Decision Approach (MDA) to reference with graded properties. Building on the comparison of the different approaches in that section, I will now proceed to an example-based evaluation, demonstrating that PRAGR as a representative of the MDA is capable of handling those aspects of gradedness which pose problems for the IDA. First, I will present an implementation of PRAGR in a world of dogs of different height and corpulence, modelling the features *height* and *corpulence* and integrating them into PRAGR. Then I will demonstrate that PRAGR is capable of (1) handling global and local context in an integrated fashion, (2) modelling the preference for properties with a higher target-distractor contrast, and (3) handling the *tall fat giraffe scenario* in a reasonable fashion.

3.3.1 Property Models

Both HEIGHT and CORPULENCE are graded properties in the sense discussed by van Deemter (2006). For each property, there is only one relevant dimension – physical height for HEIGHT and the ratio of weight to height for CORPULENCE. The main difficulty lies in the context dependence of graded properties.

According to van Deemter (2006) acceptability of graded properties depends on both local and global context. The influence of local, or situational context concerns the potential distractors present in the scene: considering a pack of small cats, even a medium-sized cat may be called large. As van Deemter (2006) further points out, whether an object is TALL or SHORT also depends on global context – the restrictions imposed by the category of an object upon the evaluation of its graded properties: "[if] Hans's and Fritz's heights are 210 and 205 cm, respectively, then it seems questionable to describe Fritz as the short man, even if Hans is the only other man in the local context" (van Deemter, 2006).

However, it makes sense to assume that for the acceptability of a graded property in a descriptive sense, i.e., the degree to which the descriptive utterance *This dog is tall.* is acceptable, only the global context is relevant – whether this particular dog is TALL *for a dog.* In contrast, the local context concerns the appropriateness of using a given term in an RE, as in *bring me the tall dog.* If the dog in question is the tallest dog in the scene by some relevant margin, the RE is appropriate, even if the dog is not TALL per se, as it has sufficient Discriminatory Power for the given purpose. Based on this reasoning, only the global context should be considered for modelling acceptability for PRAGR, while local context should be left for the PRAGR mechanism to handle in terms of appropriateness, mediated by Discriminatory Power. Thus, even without considering local context when determining acceptability values, PRAGR should use the property TALL for a given dog, even if it is fairly SHORT for a dog *per se*, as long as TALL discriminates the dog in the given context. At the same time, basing acceptability on global context allows the system to choose the more acceptable property when two

order to calculate the Acceptability of the property. A sensitivity parameter determines how fast Acceptability declines with increasing distance from the prototype. In the examples given in this section, the sensitivity of HEIGHT is set to 0.025. Feature modelling, the similarity equation, and the rationale behind parameter settings are explained in detail in Chapter 4. For the purpose of the current evaluation, it is sufficient to know that the model returns Acceptability values $\in [0, 1]$ which will be listed for each example.

Corpulence is modelled accordingly, using the weight-height ratio in kg/cm as a basis for modelling: $corp = weight/height$. In our dog world, dogs can have a corpulence between $0.1\,kg/cm$ and $1\,kg/cm$. The dog with the highest possible weight/height ratio ($1\,kg/cm$) is considered the prototype for FAT, while the dog with the lowest weight/height ratio ($0.1\,kg/cm$) is considered the prototype for SKINNY. In the examples shown here, the sensitivity of CORPULENCE is set to 2.2 (again, see Chapter 4 for a discussion of how parameters were set).

Note that the large difference in sensitivities for height and corpulence originate in the fact that they use different units of measurement. The values were set ad hoc such that all properties cover an acceptability range from 0.14 to 1.0. No claim is made here that this setting corresponds best to human judgment. Appropriate settings for sensitivity should ideally be derived from human acceptability judgments. Accordingly, the judgments of PRAGR for discriminatory power may deviate slightly from human judgment. However, these values serve well enough to demonstrate the functioning of PRAGR in principle, and to show how numerical values in relevant dimensions can be transformed to acceptability values, and how the evaluation of discriminatory power results from those acceptability values.

3.3.2 Interaction of Global and Local Context

First, I will demonstrate how the proposed model, using acceptability values based entirely on global context, nevertheless considers local context in order to determine the appropriateness of an RE. Scenario 1, as represented in Figure 3.3 and Tables 3.1 and 3.2, presents a scene with two dogs. Figure 3.3

SHORT, as can be seen in Table 3.1 – remember that dogs in this world can be from 10 *cm* to 100 *cm* tall, while the specific dogs shown here are 90.97 *cm* and 99.35 *cm* tall, respectively. Nevertheless, in order to identify dog 1, we would expect an effective REG algorithm to call it *the short dog* rather than *the tall dog*, as the latter would not allow a listener to identify the intended target object. Table 3.2 shows that the discriminatory power of the concept SHORT for dog 1 calculated by PRAGR improves chances of selecting the correct dog over chance (0.55, where chance would be at 0.5). As the PRAGR algorithm was called with $\alpha = 0$, indicating that discriminatory power alone should determine Appropriateness, PRAGR chooses the description *the short dog* for dog 1, providing an effective description of the animal.

Compare this to scenario 2, described in Figure 3.4 and Tables 3.3 and 3.4, where a much shorter dog than the one just described as *the short dog*, namely dog 1 in Figure 3.4 is described as *the tall dog*, because in this case the distractor is smaller, and *the tall dog* has a high discriminatory power: $P(o_1|D) = 0.68$. These two examples show clearly how the principle of Discriminatory Power allows the consideration of local context for REG while using Acceptability values derived solely from global context. Thus, PRAGR generates expressions which are in line with van Deemter's (2006) observations regarding local context, without requiring this to be explicitly modelled in the Acceptability of the properties. This allows a clear separation between the Semantics of a graded adjective as represented by its *acceptability* function, which should be based on global context (i.e., how tall a dog is *for a dog*), and the Appropriateness of using that adjective for *referring* to an object *in a given context*. By increasing α, the influence of Acceptability, and thereby global context, can be increased, an issue which will not be explored further at this point.

3.3.3 Target-Distractor Contrast

As discussed in Section 2.3.3, when two different graded attributes are available to identify an object, humans select the property with the largest object-distractor contrast (Hermann and Laucht, 1976). If we look at Figure 3.4,

by PRAGR for dog 1 is much higher than that of TALL (0.84 as compared to 0.68). As expected, PRAGR describes dog 1 as *the skinny dog*.

In contrast to this behaviour, as discussed in Section 3.1.4, the Global Decision Approach (GDA) does not consider differences in discriminatory power between different properties beyond checking whether each property reaches a predetermined minimal discrimination threshold. Thus, the GDA cannot achieve the subtle distinctions made by PRAGR in the scenarios described here.

3.3.4 Handling the Tall Fat Giraffe

As van Deemter (2006) argues, an REG mechanism which handles graded properties should generalise to combinations of graded adjectives such as *the tall fat giraffe* which "might describe a referent that is neither the tallest nor the fattest giraffe, as long as a combination of height and fatness singles it out" (van Deemter, 2006, p. 199). In Section 3.1.4, I argued that a Modular Decision Approach such as PRAGR handles *tall fat giraffe* cases straightforwardly, both in REG and RR, a claim I will substantiate in the following with scenarios 4, 5, and 6.

Scenario 4 (Figure 3.6, Tables 3.7 and 3.8) is a scene of four dogs. Figure 3.6 shows a visual representation and the description generated by PRAGR for each dog. Measurements and acceptability values for individual properties are given in Table 3.7, while Table 3.8 shows Acceptability and Discriminatory Power for potential descriptions for the different dogs. As in the prior scenarios, each dog was described independently by PRAGR, with $\alpha = 0$. Unlike the prior examples, descriptions of up to two properties were allowed in order to examine PRAGR's treatment of combinations of properties.

Dog 2 is neither the TALLEST nor the FATTEST dog in the scene – dog 4 is taller, while dog 3 is fatter than dog 2. To my knowledge, no empirical research on human behaviour in producing or interpreting expressions in such tall-fat-giraffe scenarios has been conducted, therefore the assumptions of which descriptions are reasonable can only be based on my own intuition. The description *the tall fat dog* is the intuitively most appealing description,

ID	Height(cm)	p(tall)	p(short)	Weight(kg)	Corpulence	p(fat)	p(skinny)	description
1	52,26	0,30	0,45	18,18	0,35	0,24	0,58	the short skinny dog
2	72,26	0,50	0,27	50,27	0,70	0,51	0,27	the fat dog
3	65,16	0,42	0,32	48,16	0,74	0,56	0,25	the short fat dog
4	94,19	0,86	0,16	46,41	0,49	0,33	0,42	the tall skinny dog

Table 3.9: PRAGR descriptions for Scene 5 with 4 dogs. Parameters: sensitivity corpulence=2.2, sensitivity height=0.025, alpha=0.0.

description	$p(D\|o_1)$	$p(D\|o_2)$	$p(D\|o_3)$	$p(D\|o_4)$	$p(o_1\|D)$	$p(o_2\|D)$	$p(o_3\|D)$	$p(o_4\|D)$
the short dog	0.45	0.27	0.32	0.15	0.37	0.23	0.27	0.13
the tall dog	0.30	0.50	0.42	0.88	0.14	0.24	0.20	0.42
the tall fat dog	0.07	0.26	0.24	0.29	0.08	0.30	0.28	0.34
the short fat dog	0.11	0.14	0.18	0.05	0.22	0.29	0.38	0.11
the fat dog	0.24	0.51	0.56	0.33	0.15	0.31	0.34	0.20
the short skinny dog	0.26	0.07	0.08	0.06	0.54	0.15	0.17	0.14
the tall skinny dog	0.18	0.13	0.10	0.37	0.22	0.17	0.13	0.47
the skinny dog	0.58	0.27	0.25	0.42	0.38	0.18	0.16	0.28

Table 3.10: Acceptability $p(D|x)$ and discriminatory power $P(x|D)$ for each description for all objects in Scene 5.

description	$p(D\|o_1)$	$p(D\|o_2)$	$p(D\|o_3)$	$p(D\|o_4)$	$p(o_1\|D)$	$p(o_2\|D)$	$p(o_3\|D)$	$p(o_4\|D)$
the short dog	0.45	0.27	0.32	0.16	0.37	0.23	0.27	0.13
the tall dog	0.30	0.50	0.42	0.86	0.15	0.24	0.20	0.41
the tall fat dog	0.07	0.26	0.42	0.28	0.07	0.25	0.41	0.28
the short fat dog	0.11	0.14	0.32	0.05	0.17	0.22	0.52	0.08
the fat dog	0.24	0.51	1.00	0.33	0.11	0.25	0.48	0.16
the short skinny dog	0.26	0.07	0.04	0.07	0.58	0.17	0.10	0.15
the tall skinny dog	0.18	0.13	0.06	0.36	0.24	0.18	0.08	0.50
the skinny dog	0.58	0.27	0.14	0.42	0.41	0.19	0.10	0.30

Table 3.12: Acceptability $p(D|x)$ and discriminatory power $P(x|D)$ for each description for all objects in Scene 6.

Cases such as described in scenarios 4, 5, and 6 cause difficulties for the IDA, as quantitative information regarding the size of contrast in each domain is lost at the combination stage. As discussed in Section 3.1.4, this can lead to faulty intermediate pruning in RR and problems handling combined descriptions such as *below and to the right of* (Gorniak and Roy, 2004, p.442).

Again, there is currently no empirical evidence on human production or interpretation of REs in such tall-fat-giraffe scenarios. However, the phenomena shown here indicate that PRAGR is capable of handling tall-fat-giraffe scenarios in an intuitively appealing way. Likewise, the results shown here point to avenues of future research, as it would be highly interesting to examine human behaviour in tall-fat-giraffe scenarios and evaluate how phenomena such as the preference for large target object-distractor contrast (Hermann and Laucht, 1976) and the pragmatic strategies found by Frank and Goodman (2012, 2014) interact in such situations. In this context, it would be desirable to compare the behaviour of PRAGR with human data for a comparable scenario in order to gain further insight into the adequacy of the account presented here.

3.4 Summary

In this chapter, I have given an overview of existing non-crisp and probabilistic approaches to reference, and have made explicit the core aims of PRAGR regarding the optimality of REs. The primary perspective from which op-

timality of REs is approached in this work is that of providing useful REs in the context of situated interaction, thus *understandability* is the perspective on optimality adopted here.

I have situated PRAGR within the MDA to reference with vague properties, an approach which calculates an integrated appropriateness score for complex descriptions, building on independent property models.

I have then presented the core mechanism of PRAGR, demonstrating how reference handling with vague properties can be approached by using a probabilistic measure of Discriminatory Power. I further identified links to property modelling, handling salience, and dealing with spatial relations.

Finally, I have provided an evaluation of PRAGR's handling of vagueness in a number of simple scenarios.

In the following chapter, I will present the general approach used for property modelling in this thesis, and the individual models used for the evaluation experiments which will be described in detail in Chapter 6.

Chapter 4

Modelling Vague Properties

In order to evaluate the performance of PRAGR beyond simple example scenarios and assess its usefulness for practical applications, it is necessary to demonstrate that a variety of properties of different domains can be integrated into a coherent functioning system. From a practical perspective, the modelling of properties for reference handling is necessary in order to be able to perform empirical evaluations of an REG or RR mechanism with realistic stimuli. However, analysing and modelling the internal structure of different conceptual domains is also of fundamental importance for the design of flexible reference handling architectures. Different conceptual domains may pose different constraints on the overall structure of the reference handling architecture. For example, the complex internal structure of the colour domain, including hierarchical relations, overlapping categories, and vague boundaries, has implications on possible restrictions of the mathematical model on the reference handling level – basing an algorithm on the assumption that all properties of a domain be mutually exclusive is not consistent with what we know about colour and may lead to fundamental problems with a reference handling approach (see also the discussion in Section 3.2.5). While a deeper analysis of these questions is beyond the scope of this thesis, the present Chapter presents a first step towards bringing together research on categorisation and reference, and working towards a view of reference informed by insights on categorisation and the internal structure of conceptual domains.

In this chapter, I will first introduce the overall framework for probabilistic property models used in this thesis – a prototype-based approach similar to Gärdenfors' (2004a) *Conceptual Spaces*, before continuing to introduce specific property models used in this thesis.

In Section 3.1.3, I have already discussed some issues which are relevant to the modelling of vague properties, in particular the distinction between degree based and delineation based approaches. As explained in Section 3.2, with PRAGR I intend to take a path which combines both perspectives, using the concept of acceptability as a bridge between subjective, graded category membership assignment, and probabilistic intersubjective estimation of communicative success. Based on this underlying assumption, I will now proceed to explain the specific approach to modelling that I use throughout this thesis.

As indicated in Section 3.1.3, there is a large amount of empirical research on categorisation which supports the so-called *prototype theory*, i.e., the assumption that categories are characterised by an idealisation of what a perfect member would be, with membership depending on overall perceived similarity to this *prototype*.

Researchers have also proposed that categories are founded on similarity to stored exemplars rather than an idealised prototype (Medin et al., 1984), and computational models based on exemplars have been proposed for a variety of domains including colour (Menegaz et al., 2007) and hairstyles (Wang et al., 2011). While there is evidence that humans use exemplars as well as idealised prototypes for categorisation (Medin et al., 1984), for practical purposes prototype representations have the advantage that (a) they are more economical, as only one similarity evaluation for the prototype needs to be made, rather than having to compare a new stimulus to a large number of exemplars, and (b) prototype representations are more flexible for the purpose of system development. It is possible to construct prototype representations entirely based on theoretical knowledge about the category at hand. Empirical data can then be used for fine-tuning parameterisation. Exemplar based models, on the other hand, require data by definition, and their quality crucially depends on the quantity and quality of available data.

For this reason, I will rely on prototype based models rather than exemplar based models in this thesis.

In order to be able to model categories based on prototypes, we require some means of determining the similarity of a given stimulus to a prototype and, based on that, derive the graded acceptability of the category for the stimulus.

4.1 Conceptual Spaces

One approach for representing categories in terms of similarities to prototypes which has received much interest is the *Conceptual Spaces* approach by Gärdenfors (2004a,b). According to Gärdenfors (2004b), conceptual domains can be represented by mixed multi-dimensional feature spaces. For example, it is possible to model colour as a combination of hue, saturation and lightness. Each dimension in a conceptual space can be represented by a numerical value. It is then possible to calculate the distance between two points as the weighted Euclidean distance, thus treating conceptual domains as geometric spaces.

Gärdenfors (2004b) follows the prototype approach to representation by suggesting to represent categories by idealised prototypical members. A prototype is formed by calculating the mean value of all category members for each dimension. According to research on the perception of similarity summarized by Gärdenfors (2004b, p. 21), the perceived similarity s_{ij} between two objects i and j can be modelled by an *exponentially decaying function* of their geometric distance $d(i, j)$. While at this point there is no agreement with regard to the exact formulation of this similarity function, Gärdenfors (2004b) proposes, amongst others, the following function:

$$s(i, j) := e^{-c \cdot d(i,j)^2} \tag{4.1}$$

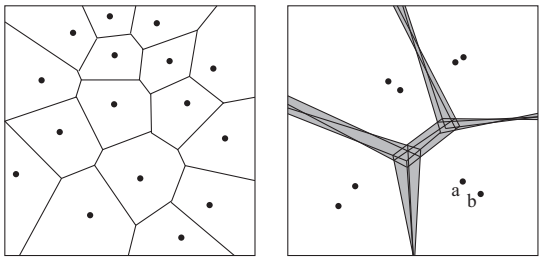
When we apply this similarity measure to categorise j based on a prototype i, c determines the specificity of the category. For example, for CRIMSON, c will be larger than for RED, leading to a faster decline of similarity,

as the category covers a smaller range of hues.

Gärdenfors (2004b) proposes that the conceptual space is divided into categories by Voronoi tessellation based on the prototypes (see Figure 4.1a), thus dividing the multi-dimensional space into convex category regions such that each individual point is assigned to the category region of the prototype it is closest to (Gärdenfors, 2004b). For example, a given triple of hue, saturation and lightness can be classified as either YELLOW or RED, depending on which colour prototype it is closest to in the colour space. The Voronoi tessellation approach effectively leads to a discretisation of conceptual space with a set of mutually exclusive, exhaustive categories, abandoning the concept of vagueness except for the boundaries of the Voronoi regions which have exactly the same distance to several prototypes (Douven et al., 2013).

In contrast, Douven et al. (2013) propose using collated Voronoi tessellation to determine regions of certain membership, vague boundary regions, and regions of non-membership, yielding a three-valued logic. In this approach, prototypes are represented as regions of different size, and a set of Voronoi tessellations is produced by performing a separate Voronoi tessellation for each combination of representatives for each prototype region (See Fig. 4.1b). When all Voronoi graphs are superimposed on each other, the region which is part of a given category for all graphs is considered the set of clear instances of this category while overlapping regions are considered to be boundary cases which do not clearly belong to the category. This approach is closely related to the so-called egg-yolk representation of spatial reasoning with indeterminate boundaries (Cohn and Gotts, 1996).

However, collated Voronoi tessellation has some counter-intuitive effects: increasing the prototypical area of a category not only increases vagueness outwards, taking in more space for potential category membership, it also increases vagueness inwards, thereby decreasing the area of clear membership. A small prototypical region, on the other hand, yields a small area of vagueness and a relatively large area of clear membership. More importantly, the main goal of this thesis is to assign acceptability values between 0 and 1 that integrate gradedness and vagueness, rather than solely the determination of



(a) (b)

Figure 4.1: Voronoi graphs for characterising category regions (Douven et al., 2013). (a) Crisp categories. (b) Superimposed Voronoi graphs and overlapping boundary regions (grey)

areas of vagueness as Douven et al. (2013).

Eyre and Lawry (2014) propose a variant of *Conceptual Spaces* which follows the probabilistic boundary interpretation while still relying on prototype-based representations. In their approach, a label is defined by a prototypical element and a distance threshold which is realised by a random variable. Thus, the probabilistic boundaries are defined by the distance to the prototype.

On the other hand, it is also possible to treat the *Conceptual Spaces* approach as a graded membership approach by interpreting the cognitively motivated similarity function presented above as a membership function. Thus, if the colour of a given object yields a value of 0.8 for its similarity to the prototype of YELLOW, this can be interpreted as a graded membership value of 0.8 for the category YELLOW. A similar approach is used by Spranger and Pauw (2012) who represent categories by prototypes and model graded category membership by an exponentially decaying membership function based on the distance to the prototype.

4.2 Modelling Vague Properties for PRAGR

In order to gain graded membership values for use in PRAGR, in this thesis I follow a methodology of determining acceptability values based directly on the concept of similarity, by applying the cognitively motivated similarity

function in Equation 4.1 to the weighted Euclidean distance between category prototypes and individual stimuli. As discussed in Section 3.2, acceptability is interpreted in a two-fold way. On the one hand, it incorporates gradedness in the sense that with increasing distance from the prototype, instances are considered to be less good examples of the category. On the other hand, it constitutes a probabilistic estimate of the likelihood that an interlocutor would accept description D for object x. Thus, the more similar an instance to the prototype, the better an example of the category it is perceived to be, and the more likely it is that an interlocutor will follow the conceptualisation of the object as a member of this category, as reflected by a higher value of acceptability.

As PRAGR is capable of selecting one of several categories of the same conceptual space, i.e., RED vs. YELLOW based on acceptability values and situational context, it is not necessary that the property model provide acceptability values of zero even for those elements which are clearly not members of the respective category. For example, whether a given colour should be called *red* or *yellow* or *crimson* will be determined by comparing the acceptability value of each of these properties for the target object with the summed acceptability value of the same properties for other objects in the scene. In this case, the acceptability value of a given ball for YELLOW does not necessarily have to be zero even if the ball in question would not be normally classified as yellow, as long as the value is sufficiently low that it will be strongly dispreferred as compared to the other options available (compare Section 3.2.5).

As mentioned above, when we apply this similarity function to categorisation of an object j based on a prototype i, the sensitivity value c determines the specificity of the category. This allows us to accommodate for complex hierarchical relations between concepts by adapting parameter c individually per concept. Therefore, hierarchical organisation is solved in a much more flexible way than allowed for by Voronoi tessellation. Throughout this thesis, I use manually set values for c which are based entirely on experimentation, comparing the acceptability values produced by a given parameter setting with my own subjective judgment. While not ideal, this is a limitation based

on pragmatic considerations, as parameter optimisation is not the focus of this thesis, and the manually determined values are sufficient for the purpose of evaluation pursued here. Values for c can easily be learnt using Machine Learning (see Section 8.3.3).

In addition to c, the weights of the different dimensions need to be determined. Here it is important to note that Gärdenfors (2004b) seems to assume that the weights of the dimensions are identical for the entire domain (e.g., the weights for hue, lightness, and saturation need only be determined once for the entire domain of colour). When looking at the colour domain, this is clearly not the case. The colour BLACK is determined first and foremost by lightness, thus the weight of lightness will be considerably larger than the weights of the other dimensions. The colour RED, on the other hand, is determined mainly by hue, allowing large variation on the dimensions of lightness and saturation, while CRIMSON covers only a narrow range of all three dimensions – hue, lightness, and saturation. Therefore, for the purpose of this thesis I assume that all parameters of the similarity function – dimension weights and sensitivity – are specific to each category, rather than general to the conceptual domain.

In a similar vein, in some cases it is practically more appealing to model different categories of one conceptual domain in separate, structurally equivalent Conceptual Spaces, rather than a single Conceptual Space, as will be discussed in Section 4.6 for the case of projective relations.

To conclude, the approach for categorisation used in this thesis is inspired by the Conceptual Spaces approach, but is not a direct application thereof. It has in common with Gärdenfors' Conceptual Spaces the geometric metaphor (weighted Euclidean distance as conceptual distance), and the modelling of categories via idealised prototypes. In contrast to Gärdenfors (2004b), I do not perform crisp categorisation with Voronoi tessellation, but rather derive acceptability values using a similarity function and I allow for individual parameterisation of different categories within a Conceptual Space.

Based on the given characterisation of the overall methodology used for modelling categories in this thesis, I will now proceed to introduce the vague property models used in this thesis. Likewise, I will introduce a number of

crisp models that are used as a baseline for comparison.

The specific features modelled in this work are chosen for the purpose of (1) demonstrating the feasibility of probabilistic property modeling for different property domains, (2) highlighting the particular challenges faced when modelling different kinds of property domains, and (3) providing a basis for the evaluation of PRAGR in Chapter 6. Table 4.1 gives an overview of all the property models presented in this thesis, their basic characteristics, and the evaluations they are used for. The following property domains will be addressed in this thesis:

- The domains HEIGHT, CORPULENCE, and SIZE are used as examples of graded properties.
- The property pair SQUARE vs. LONG(ISH) is used as an example for domains which behave asymmetrically: the property SQUARE is generally considered to be crisp, while the property LONG is considered to be graded, even though they both operate on the same domain, namely the ratio between the two relevant size dimensions (e.g. height vs. width in a 2 dimensional image).
- The colour domain is chosen as an example of a multi-dimensional feature domain with vague boundaries. Colour has been demonstrated to be a highly preferred attribute in human reference production (e.g. Pechmann, 1989; Viethen et al., 2012). Moreover, colour exhibits a number of interesting effects such as non-exclusive properties (e.g. an object can be both RED and DARK-RED) and hierarchical relationships (DARK-RED is a subtype of RED).
- SHAPE is an example of a property with a high danger of inaccuracy in robot perception, as detecting shapes from photographs of real objects is a non-trivial enterprise due to variation in size and rotation, and distortions which depend on the viewing angle, among other difficulties (Latecki et al., 2000).
- In order to demonstrate PRAGR's ability to handle relations, I have modelled simple projective terms, with one variant of the model for

Table 4.1: Overview of property models

Nr	Crisp/ Vague	Domain	Categories	Dimensions	Evaluation name	Section (link)
1	Vague	Height (global prototype)	Tall, Short	Height (cm)	Tall Fat Giraffe	3.3
2	Vague	Corpulence (global prototype)	Fat, Skinny	Weight/Height (kg/cm)	Tall Fat Giraffe	3.3
3	Vague + Crisp	Size (global prototype)	Large, Small	Area (Pixels)	Generating for humans	6.4
4	Vague + Crisp	Size (local prototype)	Large, Small	Area (Pixels)	robot-robot; understanding human REs	6.2, 6.3
5	Vague + Crisp	Simple Shape	Square, Long	ratio short side / long side	generating for humans	6.4
6	Vague + Crisp	Colour	37 simple and modified colour terms (see Table 4.2)	Hue, Lightness, Saturation	robot-robot; understanding human REs	6.2, 6.3
7	Vague + Crisp	Rainbow Colours	orange, yellow, green, turquoise, blue, purple, pink, red	Hue	generating for humans	6.4
8	Vague + Crisp	Shape	square, triangle, circle	optimal partial shape similarity (distance from minimum)	robot-robot; understanding human REs	6.2, 6.3
9	Vague + Crisp	Projective relations	left, right, in front of, behind	Centre point angular deviation, bounding box angular deviation, minimal physical distance	robot-robot	6.2
10	Vague + Crisp	Projective relations	left, right, above, below	Centre point angular deviation, bounding box angular deviation, minimal physical distance	generating for humans	6.4
11	Vague + Crisp	Projective relations	left, right, above=behind, below=in front of	Centre point angular deviation, bounding box angular deviation, minimal physical distance	understanding human REs	6.3
12	Vague + Crisp	Spatial Regions	left, right, front, back, middle	on-axis distance to extreme point	understanding human REs	6.3

horizontal projection (*left*, *right*, *in front of*, and *behind*) and another for vertical projection (*left*, *right*, *below*, and *above*).

- Projective regions were added due to their frequent occurrence in the corpus of human-produced REs collected for the evaluation of PRAGR.

4.3 Gradable Adjectives

Some issues with modelling graded adjectives were already mentioned in Section 3.3.1, and will be briefly summarised here.

I have demonstrated above that by modelling the Acceptability of graded adjectives using global context, it is possible to capture the effect of local context via the concept of Discriminatory Power in the reference handler.

An additional issue that needs to be considered in scenarios with several different types of objects is that graded adjectives such as SMALL or LARGE, are dependent on linguistic context – e.g., one might call the same animal *a large mouse* vs. *a small animal* (van Deemter, 2006). This issue is not considered in this thesis. In practice, PRAGR's mechanism of Discriminatory Power will gloss over this issue due to its robustness, in particular as the scenarios for empirical evaluation contain only geometric objects for which prototypes will not vary strongly. Gärdenfors (2004b) suggests to represent class dependent properties as subspaces within the conceptual spaces of the more general domains. Applied to PRAGR, this means using class-specific prototypes and sensitivity values and the integration of constraints into the PRAGR mechanism such that LARGE_{mouse} could only be combined with the head noun MOUSE, not with ANIMAL or ELEPHANT, an issue which is left for future work.

The domain HEIGHT is modelled only for the purpose of the dog world demonstration discussed in Section 3.3. The dog world can contain dogs between 20 *cm* and 100 *cm* in height. Therefore, the prototype of TALL is set to 100 *cm* and the prototype of SHORT to 20 *cm*. The Acceptability of TALL or SHORT is calculated by applying Equation 4.1 to the difference between the height of the target object and the prototype of the respective property.

The sensitivity parameter of HEIGHT is set to 0.025 for all examples[1].

CORPULENCE is modelled accordingly, using the weight-height ratio in kg/cm as a basis for modelling: $corp = weight/height$. The highest possible weight/height ratio in the dog world ($1\,kg/cm$) is considered the prototype for FAT, while the lowest weight/height ratio ($0.1\,kg/cm$) is considered the prototype for SKINNY. In the evaluation examples, the sensitivity of CORPULENCE is set to 2.2.

The domain SIZE is used in the empirical evaluation experiments. The relevant dimension used here is the area of the objects in pixels. While this is an exact measure in the RR experiment with 2D images, size is estimated based on a projection of the photographs of 3D scenes for the other experiments. For the generation of human-understandable utterances, a size model based on pixels is used with fixed prototypes (LARGE: 800 px, SMALL: 140 px) and $c = 0.002$ for both concepts.

For the robot-robot experiment and the understanding of human produced utterances, the local context is used for determining the prototypes. This has purely practical reasons, as the minimum and maximum values for size in the task is not known a priori due to the fact that the stimuli are photographs of real scenes rather than computer-generated scenes. While this is not ideal, PRAGR is robust enough to cope with this minor change. The prototype of LARGE is assumed to be the largest object in the scene, while the prototype of SMALL is the smallest object in the scene. Then the similarity function from Equation 4.1 is applied to the difference between the area of the target object in the image and the area of the respective prototype, using a sensitivity value of $c = 0.0004$.

Figure 4.2: HSL colour space with the labels defined by the *vague* Qualitative Colour Descriptor (*vQCD*) (Falomir et al., 2014).

4.4 Colour Model

Colour is the property domain which most naturally lends itself to the *Conceptual Spaces* approach. A range of geometric representations of colour have been proposed in the past for different purposes, and Gärdenfors (2004b) makes extensive use of colour as an example to demonstrate the merits of the Conceptual Spaces approach.

The physical properties of colour can be described either additively or subtractively (Brainard, 2003). The RGB (red – green – blue) colour model is an additive colour system which describes colours in terms of the intensity of red, blue, and green light. It is widely used in digital devices which produce images by emitting light from three very closely allocated light sources with the three basic colours. The RGB model is often represented geometrically as a cube with the three basic colours forming the three dimensions. On

[1]As mentioned above, all weights used in this thesis are determined by experimentation and subjective judgment and no claim is made as to their empirical validity. They are sufficient for the purpose of this work, but in order to obtain empirically valid weights, machine learning methods should be applied.

the other hand, CMYK (cyan – magenta – yellow – key) is a subtractive colour model which describes colours in terms of masking a white background with coloured ink (in the colours cyan, magenta, yellow, and black), thus subtracting brightness with each layer of colour applied.

While these physically motivated colour models serve well for technical purposes, Palmer (1999, p. 97) emphasises that "[t]he subjective experience of surface colour has a very different structure from that of physical light", depending also on the physiological properties of the human visual system. Building on the insight that "[a]ll the surface colours experienced by a person with normal colour vision can be described in terms of just three dimensions: 'hue', 'saturation' and 'lightness' " (Palmer, 1999, p. 97), a number of perceptually based colour systems have been devised, starting with the Munsell colour system (Brainard, 2003).

In this tradition, the HSL (hue, saturation, lightness) colour space represents each colour using the three perceptual dimensions *hue*, *saturation*, and *lightness*. The hue subspace is represented by the angle in a circle, while saturation and lightness are linear. Due to their interrelatedness, the latter two form a triangle, yielding a spindle as the conceptual space of colour, as shown in Figure 4.2 (compare Gärdenfors, 2004b). The HSL colour space is a geometric transformation of RGB colour space, i.e., it represents the same range of colours and can be directly mapped to it. HSL and related colour spaces are widely used in image processing and computer graphics, as these systems have the advantage of being directly mappable to RGB, while representing colour in a way that better resembles humans' intuitive colour perception and traditional colour theory in the domain of arts (Palmer, 1999; Sarifuddin and Missaoui, 2005). Moreover, HSL also coincides well with human colour naming strategies: the hue dimension corresponds to basic colour terms such as *red*, *pink*, or *purple* while the dimensions of saturation and lightness allow modifying colour names with labels which refer to the perceived richness and brightness of the colour, such as *pale-green* or *dark-blue* (Sarifuddin and Missaoui, 2005).

HSL colour space has been criticised as not being perceptually uniform, i.e. the difference between two colours in HSL does not correspond directly to

their perceived difference (Sarifuddin and Missaoui, 2005). As a consequence, several perceptually uniform colour spaces have been developed in order to adequately represent colour difference perception, an example of which is the CIE L*a*b* colour space. CIE L*a*b* has been applied by Bleys et al. (2009) for modelling the evolution of colour language in robot communities.

However, the question as to which colour model best represents human colour perception and naming has not yet been resolved. While affording the advantage of perceptual uniformity, the CIE L*a*b* colour space shows the undesired effect of weak hue constancy, such that changes in lightness or chroma affect the perceived hue (Sarifuddin and Missaoui, 2005). Moreover, CIE L*a*b* has been optimised to represent perceived similarity over very small distances and does not translate straightforwardly to larger distances (Menegaz et al., 2007; Seaborn et al., 2005). In an empirical evaluation, Seaborn et al. (2005) demonstrate that Hue, Saturation, Value (HSV) colour space – a model with similar properties to HSL – outperforms the supposedly better CIE L*a*b* and related colour systems at estimating perceived colour similarity over larger distances.

To conclude, HSL space has the advantages of straightforward mapping to RGB space, hue constancy, and good correspondence to colour naming, and the disadvantage of lack of perceptual uniformity. We can conclude that overall, HSL provides a reasonable framework for modelling colour perception for the purpose of colour naming (Palmer, 1999; Sarifuddin and Missaoui, 2005).

In the literature, a number of vague colour models have been defined. Menegaz et al. (2007) define and evaluate a fuzzy colour model which assigns each point in CIE L*a*b* space a graded membership value for one of 11 basic colour terms. For each of 424 OSA-UCS colour samples, the relative frequency of classification as each colour category is assumed to be the graded membership value, while for the remaining points, values are calculated by linear interpolation.

Seaborn et al. (2005) present a fuzzy colour model for 11 basic colour terms using *consensus areas* as prototypes, and deriving graded category membership of all other points by using a fuzzy C-means membership func-

tion. This function determines the graded membership value of a point in colour space by applying a parameterised similarity function (similar to the one described in Equation 4.1) to the distance of the colour point from the closest point of each prototype region (Menegaz et al., 2007).

Soto-Hidalgo et al. (2010) present a general approach for transforming crisp colour models into fuzzy colour models, based on prototypes and Voronoi tessellation.

The most extensive vague colour model thus far has been presented by McMahan and Stone (2015) who use a Bayesian framework and machine learning for learning a vague colour model with 829 English colour terms which considers both acceptability of terms for a given colour point, and global preferences for different terms.

In this thesis, I use a vague Qualitative Colour Description (vQCD) model which is derived from the crisp Qualitative Colour Description (QCD) model presented by Falomir et al. (2013, 2015) by transforming interval representations to centre and radius representations. A variant of this vague colour model has been described by Falomir et al. (2014). QCD is suitable as a basis for determining a vague colour model, as (1) it uses HSL space, the advantages of which have been explained above, (2) with altogether 37 simple and combined colour terms, it provides a wide range of colour terms which coincide well with those used by human subjects in an evaluation experiment (compare Section 6.3), (3) it allows straightforward adaptation to a vague colour model based on prototypes with only minor adjustments, as will be explained below.

In the following sections, I will describe in detail the process of deriving vQCD from QCD, and the resulting model.

4.4.1 Defining the Vague Colour Descriptor[2]

In direct analogy to the colours defined in the original QCD model (Falomir et al., 2013, 2015), the *vague* Qualitative Colour Reference System (*vQCD*) is defined as a set $\bigcup_i^n vQCD_i$ of individual families of colour name references which are interpreted over HSL space $[0°, 360°] \times [0, 100] \times [0, 100]$ in which the hue angle of 360° is identified with 0°. A single colour name reference (l, a_l) is a pair composed of label l and its according acceptability function a_l : HSL $\rightarrow [0, 1]$. The following five groups of vague colours are defined (Figure 4.2 shows how these are arranged in HSL space):

grey – the $vQCD_1$ family represents unsaturated colours using the labels *black, dark-grey, grey, light-grey,* and *white*.

rainbow – the $vQCD_2$ family represents saturated colours using the labels *orange, yellow, green, turquoise, blue, purple, pink,* and *red*

pale – the $vQCD_3$ family represents low-saturated colours using labels of type *pale-C* with C being a colour label defined in $vQCD_2$, e.g., *pale-green*.

light – the $vQCD_4$ family represents light colours using labels of type *light-C* with C being a colour label defined in $vQCD_2$, e.g., *light-green*.

dark – the $vQCD_5$ represents dark colours using labels of type *dark-C* with C being a colour label defined in $vQCD_2$, e.g., *dark-green*.

Inspired by Palmer and Schloss (2010), some equivalent colours were defined, namely: *dark-orange*≡*brown*, *dark-yellow*≡*olive*, *pale-red*≡*pastel-pink* according to the Inter-Society Color Council—National Bureau of Standards (ISCC-NBS[3]).

[2]Note that the adaptation of QCD to vQCD was joint work with Zoe Falomir, Daniel Couto Vale, Lledó Museros, and Luis Gonzalez-Abril. An earlier version of the adapted model was published in (Falomir et al., 2014)

[3]`http://tx4.us/nbs-iscc.htm` (Accessed August 2014)

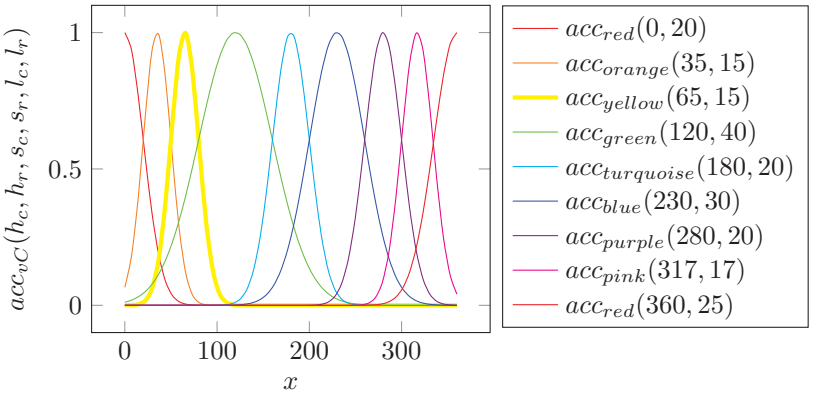


Figure 4.3: Vague colour acceptance functions shown in one dimension, the hue (h_c, h_r) dimension, for the colours of $vQCD_2$ (image adapted from Falomir et al., 2014).

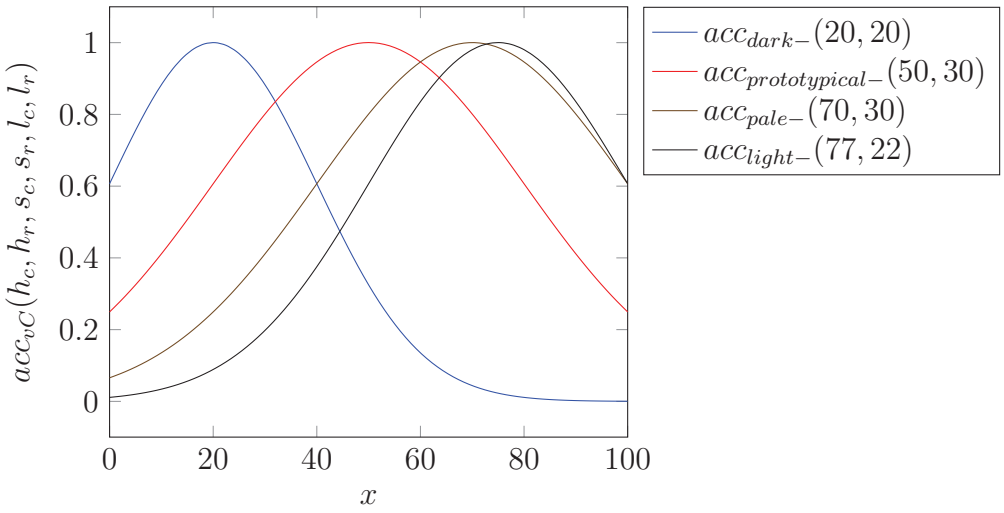


Figure 4.4: Vague colour acceptability functions shown in the lightness dimension (l_c, l_r) (image adapted from Falomir et al., 2014).

4.4.2 Vague Acceptability Functions

For each colour term l, the graded Acceptability function a_l needs to be defined based on the distance to the respective prototype. Each colour in the original crisp QCD model is defined in HSL colour space as regions $[h_0, h_1] \times [s_0, s_1] \times [l_0, l_1]$. In order to transform a crisp colour model into a vague colour model, we can make use of the fact that mathematically, each bounded interval can be equivalently described by open balls (Borelian

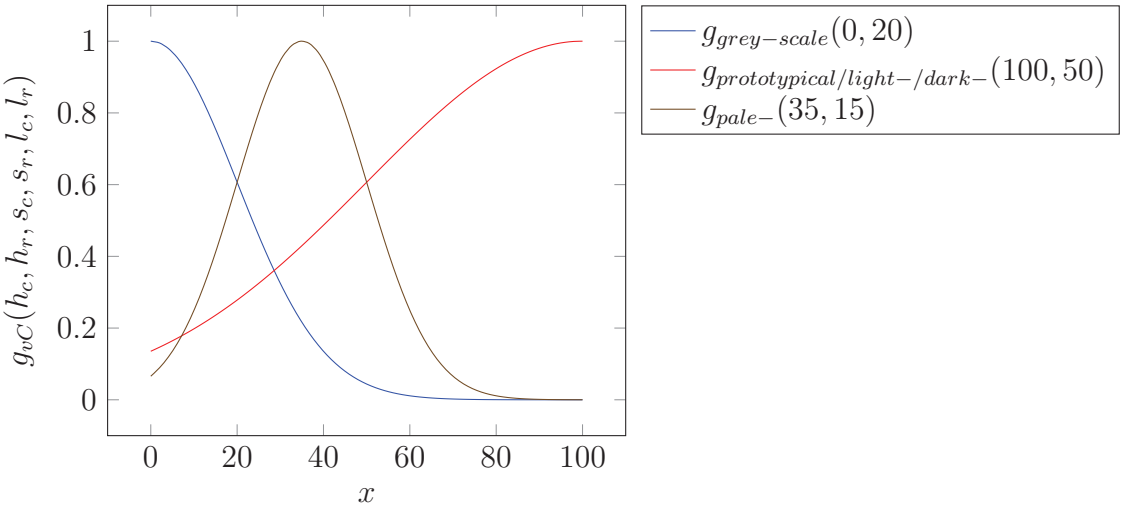


Figure 4.5: Vague colour acceptance functions shown in the saturation dimension (s_c, s_r) (image adapted from Falomir et al., 2014)

notation) $B_{h_r}(h_c) \times B_{s_r}(s_c) \times B_{l_r}(l_c)$ with h_r giving the radius in the hue dimension $\frac{1}{2}(h_1 - h_0)$, h_c giving the centre $\frac{1}{2}(h_1 + h_0)$, and analogously for saturation and lightness.

Thus, the following Radial Basis Function (RBF) determines the Acceptability for each defined colour $C = (h_c, h_r, s_c, s_r, l_c, l_r)$ as $acc_C : \text{HSL} \to [0, 1]$ by

$$acc_C(h, s, l) = e^{-\frac{1}{2}((\frac{h-h_c}{h_r})^2)+(\frac{s-s_c}{s_r})^2)+(\frac{l-l_c}{l_r})^2))} \tag{4.2}$$

where the radius in each dimension is the inverse of the sensitivity value or weight for this dimension, making this representation equivalent to the similarity function described in Equation 4.1.

Figure 4.3 presents the acceptability functions of the prototypical colours, shown in the dimension of hue according to Equation (4.2) and the parameterisation given in Table 4.2. Figures 4.4 and 4.5 show the acceptability functions on the lightness and saturation dimension, respectively. For the sake of simplicity, representations in one dimension are chosen.

Due to the spatial structure of the HSL colour space, a number of modifications need to be performed. In $vQCD_1$ (grey colours), colour terms correspond to cylinders in colour space (see Figure 4.2). Since these are inde-

pendent of hue, acc_C is rewritten as

$$acc_{(s_c,s_r,l_c,l_r)}(s,l) = e^{-\frac{1}{2}((\frac{s-s_c}{s_r})^2)+(\frac{l-l_c}{l_r})^2))}, \tag{4.3}$$

where $s_c = s_0$ and $s_r = s_1$, positioning the centre at the lower extreme of saturation. For defining the acceptability of *black*, $l_c = 0$ and $l_r = l_1$, for *white* $l_c = 100$ and $l_r = l_1 - l_0$, i.e., the prototype is positioned at the extreme lightness values. For all other colours the middle point of the cylinder axis is chosen for l_c, i.e., $l_c = \frac{l_0+l_1}{2}$, and $l_r = \frac{l_1-l_0}{2}$.

In all other cases, colour values in HSL correspond to wedges in HSL space. For all fully saturated colours ($vQCD_2$,$vQCD_4$,$vQCD_5$), the centre is positioned at the maximal saturation: $s_c = 100$ and $s_r = s_1 - s_0$. For all other cases, the barycentre of a wedge serves as centre h_c, s_c, l_l and radii h_r, s_r, l_r are determined to fit the ball into the wedge. To acknowledge wrap-around of hue space at 360° to 0°, the colour RED is associated with the maximum of two acceptability functions, one from 0° upwards and a second from 360° downwards (with the labels used here, no other colours are affected by the wrap-around).

Finally, in order to capture colour hierarchies, such as *pale-red*, *light-red*, and *dark-red* being subtypes of *red*, the acceptability of rainbow colours was broadened in the lightness and saturation dimension to increase the overlap with pale, light and dark colours (see Figures 4.4 and 4.5 and Table 4.2). Likewise, GREY overlaps LIGHT-GREY and DARK-GREY.

4.4.3 Variant: Rainbow Colour Model

Apart from the full colour model which was used for the robot-robot evaluation and for the experiment on understanding human utterances, a reduced colour model was also used for the experiment on generating human understandable REs.

For this scenario, only the rainbow colours $vQCD_2$ were used, as the scenario only contained fully saturated colours with constant lightness.

Table 4.2: Parameters of the radial basis functions (C_i, R_i) for the colour model.

	Colour Name	**H** (C_i, R_i)	**S** (C_i, R_i)	**L** (C_i, R_i)
$vQCD_1$	*black*	n.a.	(0, 20)	(0, 20)
	dark-grey			(20, 10)
	grey			(50, 20)
	light-grey			(62, 12)
	white			(100, 25)
$vQCD_2$	*red*	(0, 20) $\wedge$ (360, 25)	(100, 50)	(50, 30)
	orange	(35,15)		
	yellow	(65,15)		
	green	(120, 40)		
	turquoise	(180, 20)		
	blue	(230, 30)		
	purple	(280, 20)		
	pink	(317, 17)		
$vQCD_3$	*pale-red*, *pastel-pink*	(0, 20) $\wedge$ (360, 25)	(35, 15)	(70, 30)
	pale-orange	(35, 15)		
	pale-yellow	(65, 15)		
	pale-green	(120, 40)		
	pale-turquoise	(180, 20)		
	pale-blue	(230, 30)		
	pale-purple	(280, 20)		
	pale-pink	(317, 17)		
$vQCD_4$	*light-red*	(0, 20) $\wedge$ (360, 25)	(100, 50)	(77, 22)
	light-orange	(35, 15)		
	light-yellow	(65, 15)		
	light-green	(120, 40)		
	light-turquoise	(180, 20)		
	light-blue	(230, 30)		
	light-purple	(280, 20)		
	light-pink	(317, 17)		
$vQCD_5$	*dark-red*	(0, 20) $\wedge$ (360, 25)	(100, 50)	(20, 20)
	dark-orange, *brown*	(35, 15)		
	dark-yellow, *olive*	(65, 15)		
	dark-green	(120, 40)		
	dark-turquoise	(180, 20)		
	dark-blue	(200, 260]		
	dark-purple	(280, 20)		
	dark-pink	(317, 17)		

4.5 Contour-Based Vague Shape Model

Shape is a crucial property for human perception and reasoning. Knowing the shape of an object allows more predictions on further relevant properties of this object than knowing any other property of the object (Palmer, 1999). Unlike colour, however, shape does not lend itself straightforwardly for representation in the Conceptual Spaces framework. Gärdenfors (2004a) argues that such a representation is possible, and refers to prior approaches to represent object types in terms of hierarchical compositions of simple shapes (Marr and Nishihara, 1978). However, representing such complex information in terms of Conceptual Spaces would require at least higher order conceptual spaces (Gärdenfors, 2004a). Moreover, the approach by Marr and Nishihara (1978) is purely theoretical and does not provide a means to relate sensor-level representations of shapes as they may be retrieved by a camera to categories such as DOG, or even SQUARE.

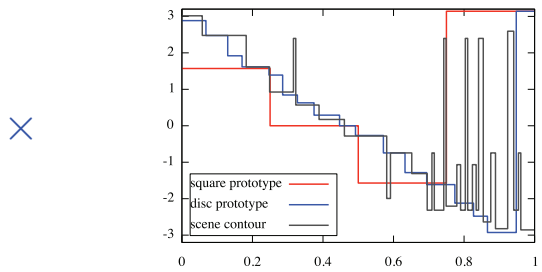
Apart from the purely theoretical shape composition approach, a number of qualitative approaches have also been proposed to measure the similarity of shapes. One kind of model describes polygons in terms of the arrangement of the individual segments to each other using qualitative spatial calculi. Such a qualitative approach to shape similarity is used, for example, by Gottfried (2008) who defines a system of *bipartite arrangements* which determines the possible relations between an oriented reference segment and another segment. A conceptual neighbourhood graph represents the conceptual similarity of the different relations. Based on an analysis of the patterns of such relations, different kinds of objects can be distinguished, and similarity between objects evaluated.

Among the quantitative approaches, Super (2004) and Ling and Jacobs (2007) measure shape similarity in terms of matching critical points on the shape contour, while others match segments of the contour (Latecki and Lakämper, 2000; Latecki et al., 2008). Mori et al. (2001) represent each shape as a sample of vertices and their position relative to each of the other vertices in the sample, allowing fast shortlisting of potentially similar shapes.

Blum (1973) presents an approach of defining shapes not by their con-

tour, but by hierarchical skeletal representations, or symmetry axes. Skeletal points are grouped according to the local variation of the radius (Blum, 1973) such that the object is the union of discs of a given radius on each point of the skeletal axis. In a further development of this approach, Siddiqi et al. (1999) propose a method for using shock graphs (a more sophisticated version of Blum's skeletal representations) for determining the similarity between shapes. They determine a number of edit operations on shock graphs, defining the similarity of two shapes by the minimal edit distance between their shock graphs. Based on the method of shock graphs, Sebastian et al. (2002) determine categories of similar shapes using a rough estimation heuristic as a means to efficiently retrieve similar shapes from a database as response to queries.

A core problem for determining shape similarity is noise. Noise can occur due to impreciseness of perception, or due to minor irregularities in shape. If noise has a large impact on the shape representation, this may deteriorate shape similarity measures dramatically. For example, Blum's approach is highly susceptible to noise, allowing small changes in contour to have a large effect on the resulting skeletal representation. Siddiqi et al. (1999) solve the issue of noise by providing a search algorithm which is capable of matching shock graphs even in the presence of random insertions and deletions of nodes.



circle	0.367
square	0.452
triangle	0.548

(a) (b) (c)

Figure 4.6: Steps of contour-based shape modeling. (a) Segmented Image with several shape contours. (b) Tangent space representation of the marked circle in overlay with prototypes for circle and square. (c) Distance to basic shape categories. (images adapted from Mast et al., 2016)

4.5.1 Optimal Partial Shape Similarity

In order to allow efficient shape-matching based on contours in the face of noise, Latecki and Lakämper (2000) propose a cognitively motivated approach for evaluating shape similarity using a combination of optimal partial shape similarity (OPS) (Latecki et al., 2005) and discrete curve evolution (DCE) (Latecki and Lakämper, 2000) which has been demonstrated to yield good results in context of the MPEG-7 shape descriptor analysis (Latecki et al., 2000).

This method reduces the noise in a contour, yielding a simplified contour representation with only few vertices. DCE is performed by iteratively removing the vertex from a contour which has the least significant contribution to the overall shape. Curve evolution stops when the significance of the least significant vertex exceeds a predefined threshold. The contribution of a vertex p to shape information, $s(p)$ is quantified by evaluating the difference between the length of the contour including p and the length of the contour without p:

$$s(p) := ||p - p^-|| + ||p - p^+|| - ||p^- - p^+||, \tag{4.4}$$

where p^+, p^- denote the neighbouring vertices, and $||\cdots||$ stands for the Euclidean norm. This process is context-sensitive in the sense that removing a vertex increases significance of neighboring vertices, thus making their deletion more unlikely.

In the next step, the distance of such a simplified contour to a shape category is determined by computing the minimal distance between detected contours and prototypes using optimal partial shape similarity (OPS) (Latecki et al., 2005). OPS can be applied as a scale- and rotation-invariant distance measure and it is determined as follows. The basic distance between two contours is computed according to Latecki and Lakämper (2000), which in case of convex shapes simply means to determine the difference in tangent space according to Arkin et al. (1991). Figure 4.6a shows a segmented image with identified shape contours, and Figure 4.6b shows the corresponding representation of the marked object and different category prototypes in tangent space. The resulting distance of the object to the basic shape categories

is shown in Figure 4.6c. Tangent space T_P represents a polygonal curve P as step function of line segment orientation $[-\pi, \pi)$ versus curvature length. After normalising contour lengths to 1 (scale-invariance), the distance of polygonal curves P, Q is given by:

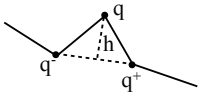
$$d(P,Q) = \int_0^1 \left(T_P(s) - T_P(s) + \Theta(P,Q)\right)^2 \mathrm{d}s, \tag{4.5}$$

where $\Theta(P,Q)$ gives the optimal alignment of the two contours' orientation and is determined by:

$$\Theta(P,Q) = \int_0^1 \left(T_P(s) - T_P(s)\right)^2 \mathrm{d}s. \tag{4.6}$$

To handle closed contours, an arbitrary point on one contour is selected as start point. Then all vertices of the other contour are tried as starting points, choosing the one that minimises shape difference. OPS is then determined by continuing curve evolution on the contour of a detected object, while keeping the prototype contour fixed, while computing the basic distance $d(P,Q)$ plus a penalty measure $r(\cdot)$ for any additional vertex removed. This process is continued until a minimum is reached which is used to derive the shape similarity measure used in this approach. The penalty measure introduced in Latecki et al. (2005) measures the distance of a vertex q to the direct line connecting its neighboring vertices q^+, q^- relative to the expected noise h_n, which in experiments is set to 2 pixels. This completes the shape distance measure d^*:

$$r(q,Q) := 0.5 \left(\frac{h}{h_n} \cdot ||q^- - q^+||\right)^2 \tag{4.7}$$



$$d^*(P,Q) := \min d(P, Q \setminus \{q_{i,1}, \ldots q_{i,m}\}) + \sum_{k=1}^{m} r(q_{i,k}, Q \setminus \{q_{i,1}, \ldots, q_{i,k-1}\}) \tag{4.8}$$

Doing so, remaining noise (see spikes in Figure 4.6b) is cancelled out in a context-sensitive manner. Finally, by considering distances between contour

Q of a scene object and all minimum distances to prototypes P_{i,i_k} of shape class $P_i = \{P_{i,1}, \ldots, P_{i,i_n}\}$ for contour Q, we obtain the desired distance measure:

$$d^*(\{P_{i,1}, \ldots, P_{i,n}\}, Q) := \min_{k=1,\ldots,n} d^*(P_{i,k}, Q) \tag{4.9}$$

4.5.2 Full Shape Model for PRAGR[4]

For the purpose of evaluating PRAGR in Chapter 6, an implementation of optimal partial shape similarity (OPS) (Latecki et al., 2005) was used in combination with DCE for simplifying shape contours (Latecki and Lakämper, 2000). This contour-based shape model is applied because of its ability to capture shape similarity in a cognitive or human-like way (Latecki and Lakämper, 2000) and to enable comparison with future studies involving complex, arbitrarily shaped objects. The distance of any given object to a shape category was determined by computing the minimal distance in tangent space between the detected contours of the object and a number of predetermined shape prototypes. For the initial noise-reduction by DCE, a threshold of 0.5 was used.

Only basic geometric shape concepts SQUARE, CIRCLE, and TRIANGLE were modelled. For SQUARE and CIRCLE, only one prototype each was used. For TRIANGLE, two prototypes were used – an acute- and an obtuse-angled triangle.

As the shape space used is very large and sparsely populated, it is possible for some objects to have a large distance to *all* prototypes, while others are close to all. However, the model is still able to distinguish different shapes with a high accuracy, as the relative distance of one object to each prototype is highly reliable. In order to capture this information, the minimal distance of the given object to any category is subtracted from all distance values, yielding a corrected distance. Eq. 4.1 is then applied to the corrected distance with a sensitivity value of $c = 50$ for SQUARE and TRIANGLE, and $c = 100$ for CIRCLE, in order to determine the acceptability of a shape. Using these

[4]The implementation of the shape model was contributed by Diedrich Wolter.

parameters yields a fairly fast decline of acceptability, reflecting the fact that for the simple shapes used here, shape perception by humans is subject to much less gradedness than the perception of, for example, colour.

4.5.3 Simple Shape Model for Generation Experiment

For the REG experiment with human participants, instead of the vague shape model, a simple shape model was used which only differentiates long and square rectangles. Using the ratio of shorter side against longer side, the prototype for SQUARE was chosen as 1.0 and for LONG as 0.25. For SQUARE, we set $c = 1.4$, for LONG $c = 0.0002$, following the intuition that acceptability for using SQUARE rapidly decreases with deviation from the prototype, while LONG is more flexible.

4.6 Projective Relations

As mentioned in Section 2.2, locative expressions – the description of a target object by its spatial relation to a reference object – are frequently used by humans for referring to objects, even in cases where they are not needed for achieving identification (Viethen and Dale, 2008). This fact is reflected in the number of publications addressing the use of spatial relations in REG (e.g. Dale and Haddock, 1991; Golland et al., 2010; Kelleher and Kruijff, 2006; Krahmer and Theune, 2002; Krahmer et al., 2003; Spranger, 2011). In the context of this thesis, I will focus on projective relations (LEFT, RIGHT, IN FRONT OF, BEHIND), as they have been widely studied both in REG and in work on cognitive modelling. In this section, I focus exclusively on modelling the Acceptability of projective relations, while the challenges the integration of (spatial) relations poses for the REG mechanism will be discussed in Chapter 5.

In the most simple sense, projective relations can be understood by imposing a co-ordinate system onto the reference object, with the direction of the main axes determined by some origin (Carlson et al., 2003; Moratz and Tenbrink, 2006). The prototypical meanings of the projective terms can then

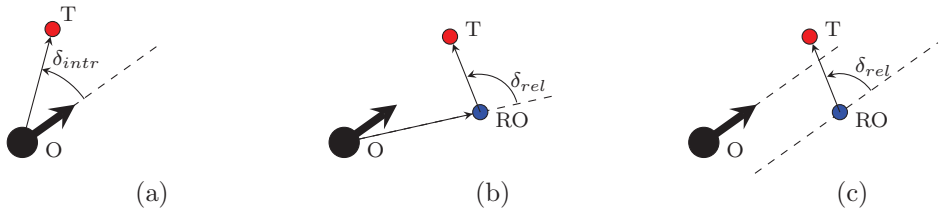


Figure 4.7: Spatial reference frames determining the angle δ in dependence of a basic reference direction. (a) Intrinsic reference frame. (b) Relative reference frame. (c) Projected relative reference frame.

be determined by the axes of this coordinate system, with graded acceptability scores depending on the deviation from the prototypical axis (Gapp, 1995b; Moratz and Tenbrink, 2006).

However, there are some complications which will be discussed in the following. Firstly, depending on the origin used to project the co-ordinate system onto the reference object, a number of different reference frames are possible. Secondly, a number of factors influence the acceptability of using a projective term, especially when one considers large reference objects.

4.6.1 Frames of Reference

Projective relations can be modelled using different frames of reference which impose a co-ordinate system onto the reference object from which the direction of the target object may be inferred (Hermann, 1990; Levinson, 1996). In the absolute reference frame, the reference direction is determined by some external bearings, such as the cardinal directions. An intrinsic reference frame (see Figure 4.7a) is given when the origin and the relatum coincide in the same object, and the reference direction is given by the intrinsic orientation of the reference object, as in *the cup is to my right.* In a relative reference frame (see Figure 4.7b), the reference direction is defined by the line between the origin (usually one of the interactants) and the relatum, as in *the cup is to the left of the bottle (seen from my perspective)* (Hermann, 1990; Levinson, 1996; Moratz and Tenbrink, 2006).

Depending on the reference frame used, a projective term may denote

entirely different regions in space, leading to ambiguity. Despite this fact, humans use reference frames flexibly and often without overt clues (Schober, 1993). The conditions and implications of the use of different reference frames have been extensively studied, both from an empirical and a computational perspective (e.g. Carlson-Radvansky and Irwin, 1993; Carlson-Radvansky and Logan, 1997; Hermann, 1990; Johannsen and De Ruiter, 2013; Schober, 1993).

In this thesis, the issue of reference frame ambiguity is not covered. The intrinsic and the relative reference frame will be used in the example-based evaluation of REG with spatial relations in Chapter 5. For the evaluation of robot-robot and human-machine interaction in Section 6, a slightly modified version of the relative reference frame will be considered. In this modified relative reference frame, the speaker projects their own intrinsic direction onto the reference object. Figure 4.7c shows how the main front-back axis is transferred from speaker O to a reference object RO, thus allowing the determination of the direction of target object T with δ indicating the deviance from the prototypical front axis. This deviates from relative reference frame discussed above where a line from the viewer to the reference object determines the direction (see Figure 4.7b). For the purpose of the experiment, it is further assumed that the listener always adapts to the speaker's perspective (see Chapter 6).

4.6.2 Graded Acceptability

While it seems that humans are capable of forming direction concepts based on either boundaries or prototypes (Crawford et al., 2000; Huttenlocher et al., 1991; Klippel and Montello, 2007; Mast et al., 2014b), there is agreement within the realm of psycholinguistics that basic projective terms such as *left* or *right* have regions of higher or lower acceptability (Carlson-Radvansky and Irwin, 1993; Gapp, 1995b; Hayward and Tarr, 1995; Logan and Sadler, 1996; Vorwerg and Tenbrink, 2007; Zimmer et al., 1998). The prototypical meanings of the projective terms are defined by the axes of the imposed coordinate system, while acceptability decreases with increasing angular deviation from

these prototypes.

Based on experiments using drawing tasks and acceptability judgments, Logan and Sadler (1996) propose modelling spatial relations in terms of spatial templates which discriminate three main regions of acceptability: good, acceptable, and unacceptable. Their results show that the templates for all projective terms are highly similar, differing only in orientation. Gapp (1995b) shows with acceptability judgment experiments that spatial relations have a broad range of acceptability which is highest for objects on the prototypical axis, and gradually decreases with increasing distance from this axis. In contrast, in a production experiment by Zimmer et al. (1998), participants only used simple projective terms such as *right* if the target was almost exactly on the axis. Otherwise they used combined terms such as *bottom right*. It must be noted, however, that in this study the listener could not see the target object, but had to search for it by moving the mouse over a blank white space, making preciseness in determining spatial relations highly relevant. Vorwerg and Tenbrink (2007) show that when responding to *where* tasks, humans tend to provide more detailed spatial descriptions than when responding to *which* tasks, indicating that a larger acceptability range for simple projective terms can be expected in *which* tasks, as compared to *where* tasks. Moreover, the forced choice due to the production task, as compared to acceptability judgments, means that the results by Zimmer et al. (1998) cannot necessarily be interpreted as graded acceptability values.

Moratz and Tenbrink (2006) present an empirically evaluated model of projective terms for human-robot interaction which yields similar results as Gapp (1995b): a broad acceptance area for spatial relations with a gradual decrease in acceptability. While Gapp (1995b) found a linear decrease of acceptability with increasing angular distance from the prototype, Moratz and Tenbrink (2006) use a smoothing function based on the cosine.

4.6.3 Spatial Relations and Gradedness in Reference

In reference, making use of the inherently graded nature of projective terms is particularly important due to a number of reasons. On the one hand, it is

often the case that a distractor stands in the same relation to the reference object as the target, and yet the relation is sufficient to discriminate the target due to a higher prototypicality of the relation to the target object, as compared to the distractor. On the other hand, humans do prefer reference objects that are in a *good* (i.e. prototypical) relation to the target (Carlson and Hill, 2009). This interaction is hard to model with crisp properties. Very broadly defined categories will lead to an undesirable increase in use of marginally acceptable relations, while narrowly defined categories may lead to unnecessarily long descriptions or inability to find a distinguishing description, as marginal cases cannot be used when appropriate.

Using the gradedness for informing reference object selection and projective term assignment allows making decisions based on subtle interactions between these phenomena, e.g., preferring more prototypical relations if they are sufficiently discriminating while using a less acceptable relation if no better discriminating description can be found.

4.6.4 Distance

While Logan and Sadler (1996) did not find any impact of the distance between locatum and relatum on the acceptability of projective terms, this aspect has received attention from researchers focusing on the task of reference object selection, i.e., selecting for a given target which object should be used as a reference object in order to describe its location. Gapp (1996, 1995a) argues that in order for an object to be a suitable reference object for locating a given target object, it needs to be close to it. Likewise, Barclay and Galton (2008) list closeness between reference object and target as one factor influencing search space reduction, a key factor for the selection of reference objects in their model. One might argue that distance is then merely a factor influencing the choice of a reference object, rather than one impacting the acceptability of a projective term. However, it must be noted that the experiment by Logan and Sadler (1996) contained only very simple, abstract scenes, and Carlson and Covey (2005) show that verbalisation of projective relations between objects implies a certain spatial closeness, indicating

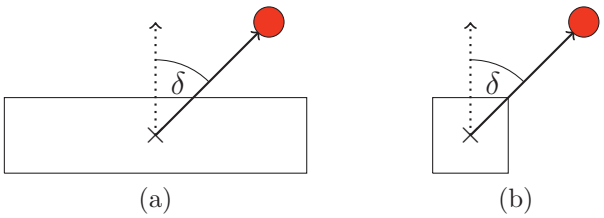


Figure 4.8: Effect of reference object size on the perception of the acceptability of ABOVE, given identical Centre of Mass angular deviation from the prototype. (a) Large reference object. (b) Small reference object.

that with increasing distance between two objects, the acceptability of any projective term will decrease.

4.6.5 Handling Large Reference Objects

The models presented so far are appealing due to their simplicity. However, as they represent both target and reference object as points, they cause some problems for situations with large reference objects. Figure 4.8 shows two scenes with the same angle between the target and reference object, if they are represented by their centroids. Due to the larger extension of the reference object in Figure 4.8a, as compared to Figure 4.8b, one would intuitively attribute a higher acceptability for the relation ABOVE in Figure 4.8a than in Figure 4.8b. However, this effect is not captured by models which use point representations of the reference object.

Using the relation ABOVE as an example, Regier and Carlson (2001) discuss a number of models that take into account the extension of the reference object while relying on point representations of the target:

- *The Bounding Box Model* is based on the intuitive assumption that, in order for the relation ABOVE to hold, the target object should be higher than the highest point of the reference object (vertical constraint), and between its leftmost and rightmost points (horizontal constraint), as shown in Figure 4.9a. According to the Bounding Box Model, if the target object is placed within this region – coloured in grey in the figure, the relation is considered acceptable. Otherwise, it is considered

not acceptable. In order to avoid crispness, Regier and Carlson (2001) apply a sigmoid smoothing function, yielding a model which returns values close to 1 if the target object is clearly within the acceptable area, and values close to 0 if it is clearly outside of this area. In the boundary region, intermediate acceptability values will be received. It is important to note that the bounding box model evaluates metric rather than angular deviation from the acceptability area. As shown in Figure 4.9b, two target objects which have the same horizontal deviation from the acceptable area would receive the same values, even if their angular deviation differs.

- *The Proximal and Centre of Mass Model* is an extension of the angular deviation approach discussed above for handling larger two-dimensional reference objects. The model combines a linear decrease for the first 72 degrees of deviation with a sigmoid decrease for deviations closer to 90 degrees. In order to account for large reference objects, Regier and Carlson (2001) form a weighted linear combination of the centre of mass orientation and the proximal orientation, where the centre of mass orientation is the angle of the line between the centre of mass of the reference object and target object, and the proximal orientation is the angle of the line between the point of the reference object which is closest to the target object, and the point-like target object.

- *The Attention Vector Sum Model* is based on psychological findings regarding attention and representation of direction: Logan (1994; 1995) shows that spatial relations are not preattentively available and thus require attention in order to be perceived. Further, Regier and Carlson (2001) base their model on the finding that several neural subsystems represent an overall direction as the vector sum of its constituent directions. Consequently, the Attention Vector Sum Model (AVS) represents directions as the weighted sum of a set of vectors distributed over the entire surface of the reference object, pointing towards the target. Vectors are weighted according to their proximity to the attentional focus which is assumed to be that part of the reference object which is

(in the case of ABOVE) vertically aligned with the target, or closest to that relation. The angle of the resulting vector sum is then compared to the prototype direction. Due to its structure, the AVS incorporates proximal and centre of mass orientation. Moreover, Regier and Carlson (2001) integrate the vertical component of the bounding box model into the model, multiplying its outcome with the outcome of the core AVS.

In evaluation experiments, Regier and Carlson (2001) demonstrate the superiority of the AVS for modelling human perception of the projective term *above*, compared to the other approaches. Their results further show that centre of mass orientation, proximal orientation, and the vertical component of the bounding box model (grazing line) are relevant components of the perception of projective relations. Moreover, they show that with increasing distance of the target from the reference object, the centre of mass orientation gains more relevance, an effect which is also captured by the AVS.

However, the AVS was primarily designed as a model of human perception, not as a component for applied systems. The model necessitates calculating vectors from all points of the reference object, thus requiring detailed shape information for the reference object. Such detailed information may not always be reliably available in situated human-machine interaction. In situated systems which rely on sensor information for situation knowledge, shape information may not be sufficiently precise to reasonably use such detailed information.

4.6.6 Dimensions for Modelling Projective Terms

The spatial models used for evaluation in this thesis take up the general insights of the AVS regarding relevant influence factors for the acceptability of projective terms with large reference objects while relying on simple, more readily available shape information and using a Conceptual Spaces approach. The following dimensions are used for modelling projective terms in this thesis:

Centre Point Angular Deviation (CP), which determines the angular deviation of the prototype of the given projective term from the line which

are formulated not in terms of one universal measure for which the prototype of each property is positioned at different points (e.g. angle from $0 - 360°$ with prototype of front at $0°$), but rather directly in terms of deviation from the respective category prototype. Thus, rather than modelling all projective terms in one Conceptual Space, as suggested by Gärdenfors (2004b), the concepts are modelled in separate, but structurally equivalent Conceptual Spaces. This is however merely a notational issue used to simplify implementation. The same information could be equally well represented in terms of a single Conceptual Space.

Based on these three dimensions, the weighted Euclidean distance to the category prototype is calculated for a given target object/reference object configuration, and the function from Equation (4.1) is applied. To summarise, the similarity to the prototype, and thus acceptability value for a given projective term can be determined as follows:

$$s(i,j) := e^{-c\cdot\left((w_{CP}\cdot CP)^2+(w_{BB}\cdot BB)^2+(w_{PD}\cdot PD)^2\right)} \tag{4.10}$$

Based on equation 4.6.6, three models of projective relations were created.

- A horizontal model of projective relations LEFT, RIGHT, IN FRONT OF, and BEHIND was used for the evaluation of robot-robot communication (model 9 in Table 4.1). For this model, the weights $W = w_{CP}, w_{BB}, w_{PD}$ were set to $w_{CP} = 0.009, w_{BB} = 0.6, w_{PD} = 0.045$.

- A vertical model of projective relations LEFT, RIGHT, BELOW, and ABOVE was used for the evaluation of robot-produced descriptions by humans (model 10 in Table 4.1). For this model, weights were set to $w_{CP} = 0.014, w_{BB} = 0.015, w_{PD} = 0.76$.

- Finally, for understanding human-produced utterances, a mixed horizontal/vertical model was created with the projective relations LEFT, RIGHT, IN FRONT OF = BELOW, and BEHIND = ABOVE (model 11 in Table 4.1), assuming identity between the front-back axis and the vertical axis. For this model, the same weights as in model 9 were used:

$$w_{CP} = 0.009, w_{BB} = 0.6, w_{PD} = 0.045.$$

For all evaluation studies, the sensitivity parameter c of the similarity function was set to $c = 1$.

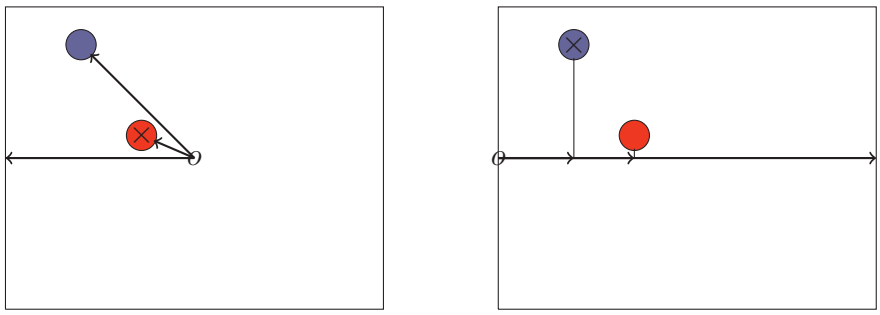
4.7 Spatial Region Model

An additional spatial model was used for evaluating understanding of human descriptions, as the human subjects frequently described the location of objects within the scene, e.g. *the large circle on the left in the middle*.

While these relations are classified as internal projection by Hois et al. (2009), their structure is different than that of external projections.

Firstly, due to the fact that the reference object is always the scene as a whole, they need not be modelled as relations, but can be treated as simple properties. Secondly, while they are also concerned with directionality, they are usually not modelled using angular deviation from the prototype direction, but rather by evaluating the position along the prototypical axis (Gorniak and Roy, 2004). The closer an object is to the extreme point on the directional axis (e.g., the leftmost point on the x-axis), the higher its acceptability for the respective direction. Figure 4.10a shows the angular deviation measure used for external projections, while Figure 4.10b depicts the measure of distance from extreme point used for modelling internal projections.

Like external projections, internal projections depend on a reference frame which determines the directionality of the prototypical axes. For the sake of this thesis, a fixed reference frame is assumed which is determined by the perspective of the speaker. The prototype of each projective location (LEFT, RIGHT, UPPER, LOWER) is set locally to the outermost object point in the respective dimension. For example, the rightmost point of the rightmost object in the scene is considered the prototype of RIGHT for this scene. Acceptability is calculated by applying Equation (4.1) to the minimal distance of an object from the prototype, considering only the relevant dimension (horizontal for LEFT and RIGHT, vertical for UPPER and LOWER). UPPER and LOWER are handled analogously. Finally, MIDDLE is also modelled as a



(a) (b)

Figure 4.10: External projection vs. internal projection. The object with higher acceptability for LEFT is marked with ×. (a) External projection determined by angular deviation from prototypical axis. (b) Internal projection determined by distance from extreme point along prototypical axis.

non-projective regional location. For MIDDLE, the model uses the centre of the extreme points on both axes as the prototype. Acceptability is calculated by applying Equation (4.1) to the Euclidean distance of the centre point of the object to this prototype. C is set to 0.0001 for all concepts.

4.8 Crisp Models for Evaluation

For the purpose of evaluating PRAGR using vague properties against PRAGR using crisp properties, for any given vague model, a corresponding crisp model can be defined based on the idea of Voronoi tessellation suggested by Gärdenfors (2004b). By performing the Voronoi tessellation on the acceptability values rather than the distance, it is possible to perform Voronoi tessellation even for domains whose properties are modelled in separate, equivalent Conceptual Spaces. Each object is assigned to that category of the respective domain for which it has the highest acceptability. The Acceptability for the assigned category is set to 1, while the Acceptability for all other categories of the domain is set to 0.

4.9 Summary

In this chapter, I introduced the framework for probabilistic property models used in this thesis – a prototype-based approach similar to Gärdenfors' (2004a) *Conceptual Spaces*, and introduced a number of specific property models which will be used in the remainder of this thesis. In particular, I introduced property models for gradable adjectives, colour, shape, projective relations, and spatial regions, discussing the choice of dimensions used for modelling, and providing variants of the models which are adapted to different application scenarios.

Chapter 5

Challenges: Spatial Relations in REG

Integrating spatial relations into REG algorithms poses a number of challenges. As spatial relations are at least binary rather than unary relations, they require the use of a second object, the reference object, in order to describe the first. This requires determining which object is suitable as a reference object. While research exists on reference object selection, as well as on REG with relations, it is by no means clear how the task of reference object selection can be integrated into REG. Krahmer and van Deemter (2012, p. 184) conclude: "On balance, it appears that the place of relations in reference is only partly understood, with much of the iceberg still under water."

We can determine two core problems of integrating spatial relations into REG. Firstly, it needs to be determined how the identifiability of the reference object can be factored into the evaluation of the optimality of a whole complex description, and whether and how the insights from research on reference object selection can be considered in REG.

Secondly, including relations into REG dramatically increases the problem of combinatory explosion, giving more urgency to the need for efficient search algorithms. Relations further pose problems for some algorithms which were not originally designed with relations in mind, such as *forced*

incrementality or infinite recursion.

In the following, I will discuss the core challenges faced when integrating spatial relations into an REG algorithm and demonstrate how PRAGR can tackle some of these issues in keeping with the approach to vagueness taken here. In Section 5.1, I will discuss the challenge of reference object selection, in particular the following aspects: (1) how to integrate the identifiability of the reference object into an overall evaluation of Discriminatory Power, and (2) the integration of salience into the REG mechanism. In Section 5.2, I will then proceed to discuss the search problem as it is posed by combining gradedness and spatial relations in a single REG mechanism and present the search algorithm used to tackle this problem.

5.1 Reference Object Selection and Optimality

When producing a complex RE including a relation, a speaker needs to select a reference object with respect to which the target object can then be described. In research on REG, reference object selection has thus far received little attention, with REG algorithms usually considering fairly simple scenes and relying on straightforward approximations (e.g., Kelleher and Kruijff, 2005; Krahmer et al., 2003). However, research on reference object selection has identified a large number of factors which influence the appropriateness of using a given reference object in order to locate a target object. While this research has thus far focused on locative expressions where the location of a known target object is described with respect to a reference object (Barclay and Galton, 2008), it is reasonable to assume that similar factors govern reference object selection for REG.

In the following, I will give a brief overview of the research on reference object selection, before discussing how the factors identified in this research may apply to REG, and how they can be incorporated into an REG mechanism which handles vague properties.

5.1.1 Research on Reference Object Selection

From the perspective of Cognitive Linguistics, Talmy (2000) describes the relationship between reference object and target object as inherently asymmetrical. The typical reference object differs from the target object in that it is less mobile, larger, treated as geometrically more complex, accessed earlier within the scene or in memory, less relevant, more immediately perceivable, backgrounded once the target object is perceived, and more independent. This asymmetry is caused by the role of the reference object which is to aid the listener at identifying the target object. If identifying the reference object were harder than identifying the target object, the reference object would hardly be able to serve this purpose.

Properties of reference objects have also been analysed in the context of wayfinding where landmarks play a crucial role for providing route instructions (Burnett et al., 2001).

The most notable models of reference object selection are the 8-factor model provided by Gapp (1995a, 1996), and the hierarchical influence model by Barclay and Galton (2008, 2013). As the model by Barclay and Galton (2008, 2013) provides a useful hierarchical grouping of factors, I will use this model in order to identify the most relevant influence factors, indicating the relation to factors proposed by Gapp (1995a, 1996) where appropriate.

In their hierarchical influence model for reference object selection, Barclay and Galton (2008) identify the main factors *locatability of the reference object*, *search-space optimisation* and *communication cost*, which are in turn composed of several factors.

Locatability of the Reference Object

According to Barclay and Galton (2008), speakers prefer reference objects which can be easily located in the scene. A reference object is easily locatable if the listener has the required knowledge to identify the reference object based on the description, and the reference object is a recognisable member of the category or categories used to describe it. These factors correspond roughly to Gapp's (1995a) notion of referentiality: if the candidate refer-

ence object cannot be reasonably described, it cannot function as a reference object. Further, Barclay and Galton (2008) state that the reference object should be visible, and persist for the duration of its use as a reference object. Visibility may depend on variables such as object size, obscurance, visual contrast, and is strongly related to the notion of visual salience which is also proposed as a factor by Gapp (1995a) – the phenomenon of some stimuli 'popping out' in a visual scene and therefore being more easily accessible. Persistence is related to Gapp's (1995a) factor of stability which states that humans prefer using stable (i.e., non-moving) reference objects to describe the location of other objects. In addition, Gapp (1995a) mentions two factors relevant for locatability of the reference object which are not mentioned by Barclay and Galton (2008), namely linguistic salience (or prior mention of the reference object) and prior knowledge of the reference object by the listener – indicating that knowledge of the reference object is part of the common ground between speaker and listener.

Search-Space Optimisation

The criterion of search space optimisation requires that the reference object, in combination with the spatial relation used for the locative expression, should optimise the region in which the listener will search for the target object. As Barclay and Galton (2008) point out, the size of the search space implied by a spatial relation depends on the size of the reference object. A large reference object implies a large search space, while a smaller reference object implies a smaller search space. Or, in the words of Miller and Johnson-Laird (1976) "It would be unusual to say that the ashtray is by the town-hall". Thus, for the selection of an optimal reference object, a balance between search space optimisation and locatability of the reference object is required. Further, Barclay and Galton (2008) note that if the target object is very small, and the distance of the listener to the target object is very large, no single reference object may suffice to adequately describe the target object. In such cases, a chain of locative expressions may be needed.

Search-space optimisation is further influenced by the position of the tar-

get object relative to the reference object: if the target object is close to the reference object, or placed on a central directionality axis according to some frame of reference, this reduces the effort for search. Empirical evidence supports this claim. For example, in a production experiment by Carlson and Hill (2009), participants preferred reference objects which were in a prototypical relation to the target, even if they were slightly less salient. In a study by Vorwerg and Tenbrink (2007), humans producing object descriptions answering a *where* question used modifiers or complex projective relations whenever the target object deviated from the prototypical axis. In a production experiment by Zimmer et al. (1998), speakers locating a target object which was invisible to the listener used simple projective terms only when the target object was very close to the prototypical axis.

Finally, the specific relation in which the target object stands to the reference object is also relevant. A number of studies show that vertical relations such as ON or ABOVE are preferred to horizontal relations such as LEFT (Bryant and Wright, 1999; Franklin and Tversky, 1990; Plumert et al., 1995; Viethen, 2011).

Regarding search space optimisation, Gapp (1995a) mainly considers distance between reference object and target object. He also mentions that objects which stand in a functional relationship with the target object are more suitable as reference object due to the fact that functionally related objects are more likely to be perceived and processed together (Hirtle and Heidorn, 1993; Hirtle and Jonides, 1985; Tulving, 1962), thus easing search.

Communication Cost

In line with the Gricean Maxim of quantity, Barclay and Galton (2008) include communication cost as a factor to account for the fact that shorter and structurally simpler REs require less processing cost (both for the speaker and the listener) than longer and more complex REs. This is relevant for reference object selection, as the choice of reference object may impact the length and complexity of the resulting description due to both the complexity of the relation between reference object and target object, and the necessity

for producing an appropriate RE for the reference object itself which may require lengthy descriptions for some reference objects.

Empirical Evaluation of Reference Object Selection

Barclay (2010) and Barclay and Galton (2013) operationalise several of these influence factors and use Bayesian Networks to learn models of up to 4 factors for generating locative expressions for objects in 3D images. They find that a combination of proximal distance between reference object and target object, convex hull volume of the reference object, and the ratio between the minimum and maximum dimension of the potential reference object best predicts human behaviour. The model matches one of the top three human choices of reference object in 73.5 % of cases.

In a reference production study using highly cluttered scenes, Clarke et al. (2013) show that humans prefer reference objects which are large, visually salient, and in close proximity to the target, thus providing evidence that both locatability of the reference object and search space optimisation are relevant factors in reference object selection.

5.1.2 Reference Object Selection and REG

Like most work on reference object selection, the work by Barclay and Galton (2008) focuses on *where* questions in which the position of a (usually already known) target object needs to be specified. This contrasts with REG which is typically concerned with *which* questions, i.e., the unique identification of a formerly unknown target (Krahmer and van Deemter, 2012).

It is an open question whether REG with relations should be considered as exclusively answering a *which* question, i.e., uniquely identifying the target from a given set of distractors, as the standard definition of REG implies, or whether it also serves a *where* question to a certain extent, i.e., specifying the location of the target object, as suggested by Clarke et al.'s (2013) definition of REG as the inverse of visual search. According to this definition, the task of REG is generating a description which will "allow somebody else to quickly and accurately locate the target" (Clarke et al., 2013, p. 1).

The conceptualisation of REG with spatial relations as a *which* vs. *where* task has implications for the definition of optimality in the case of REs with spatial relations. Vorwerg and Tenbrink (2007) show in object description experiments that when confronted with *where* questions, humans provide more detailed spatial descriptions than in *which* tasks, indicating that the tasks are understood differently. In particular, participants used significantly more precisifications and compound projective terms in the *where* task which indicates that search space optimisation plays a more important role in *where* tasks than in *which* tasks. On the other hand, in *which* tasks, the overall configuration of objects was more relevant, with participants preferably choosing properties for an object which discriminated it from other objects in the context (Tenbrink, 2005).

A crucial difference between the two interpretations is that if reference is exclusively concerned with *which* questions, the reference object need not be uniquely described independently of the target. In this case, it would be sufficient that an RE uniquely identifies the target object. If, on the other hand, REG using relations also needs to answer a *where* question, guiding visual search, it is crucial that the reference object is identified uniquely independently of the target, as search for the target object is contingent on successful identification of the reference object.

In the following, I will discuss the most relevant conditions for uniqueness of complex descriptions involving relations in the REG literature, and where they fall in terms of the *which* vs. *where* distinction.

Independently Unique Reference Object Condition

Figure 5.1 serves to illustrate this difference. For Figure 5.1a, the description *the circle to the left of the square* would definitely be considered discriminating. This scene fulfils the strongest requirements for a distinguishing description in REG with relations, namely that the reference object needs to be uniquely discriminated by the sub-description of which it is the head noun, in this case *the square*. This approach is used, for example, by Roy (2002). In his work, if the Discriminatory Power of a simple description of the target

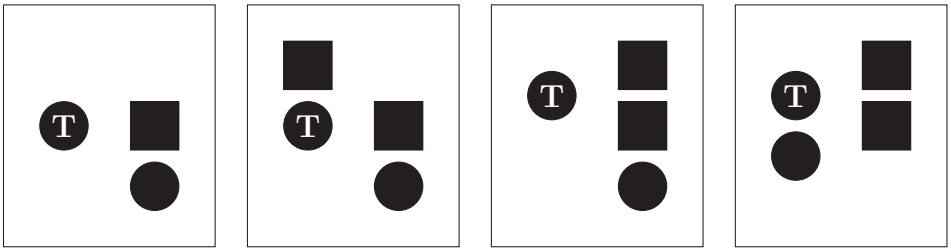


Figure 5.1: Consequences of different definitions of distinguishing descriptions for interpreting the utterance *the circle to the left of the square*. (a) Definitely discriminating. (b) Discriminating for subgraph matching. (c) Discriminating for indefinite interpretation. (d) Target preferred for total probability.

is below a given threshold, the system checks for each candidate reference object whether it can be independently described by a combination of unary relations with a Discriminatory Power above the threshold. If this is the case, this object may be selected as a reference object. This approach is in line with Clarke et al.'s (2013) interpretation of REG as guiding visual search, as it implies that the reference object needs to be identified independently of the target object in order to be useful as an reference object.

Subgraph Matching Condition

In Figure 5.1b, on the other hand, the complete description *the circle to the left of the square* uniquely discriminates the target object (marked with a T) and the reference object despite the fact that the reference object is not discriminated by the sub-description *the square* independently of the whole utterance, a phenomenon first noted by Dale and Haddock (1991). Thus, the identification of the target object and the reference object are mutually dependent. The complete description *the circle to the left of the square* can only refer to the combination of the target object and the intended reference object, as no other configuration of objects fits the description as a whole.

While the system proposed by Roy (2002) would not consider such descriptions, a number of approaches have been published which use this cri-

terion (Dale and Haddock, 1991; Kelleher and Costello, 2009; Krahmer et al., 2003). Dale and Haddock (1991) and Kelleher and Costello (2009), in extensions of the Greedy Heuristic Algorithm (GH) and the Incremental Algorithm (IA), respectively, interpret complex descriptions in the form of constraints which operate over the full scope of the RE. For example the utterance *the bowl on the table* is considered discriminating, if there is exactly one bowl which is on exactly one table, irrespective of whether there are other tables in the scene.

The same view is expressed in slightly different terms by Krahmer et al. (2003) who view REG in terms of subgraph matching. Scenes are represented as graphs with objects as nodes and properties and relations as edges. A distinguishing description is a subgraph which fits the scene graph only once.

According to this view of REG with relations, the description *the circle to the left of the square* would be considered discriminating in Figure 5.1b, as there is only one configuration of circle and square such that the circle is to the left of the square. The description would not be considered discriminating in Figures 5.1c or 5.1d, as in those cases two configurations of square and circle fit the description even though each of these configurations includes the target object.

This approach to REG with relations does not consider the issue of visual search, as identifying the reference object may require simultaneous identification of the target object. Thus it treats REG with relations entirely as a *which* question, focusing on logical discrimination rather than visual search.

Total Probability Approach

In Figure 5.1c, despite the fact that as humans we are able to identify the intended target for the utterance *the circle to the left of the square*, it is not even clear from the entire description which of the two squares in the scene should be considered the reference object. As there are two underlying target object/reference object configurations, the subgraph matching criterion would not consider the description to be distinguishing for Figure 5.1c.

In Figure 5.1d, even if one does not require the reference object to be

identified uniquely, the utterance *the circle to the left of the square* is ambiguous, though one might prefer the target over the distractor due to the fact that it is in the required relation to two potential reference objects, while the distractor is in the desired relation to only one potential reference object.

This intuition is supported by approaches which use the law of total probability in order to evaluate $P(x)$ in relation to potential reference objects y_i:

$$P(x) = \sum_{i=1}^{N} P(x|y_i) \times P(y_i) \tag{5.1}$$

This approach is followed by Engonopoulos and Koller (2014) and Golland et al. (2010). In both cases, $P(y_i)$ is calculated based only on the sub-description given for y_i, i.e., *the square* in our example. Thus, if all possible reference objects which fit this sub-description allow the unique identification of the target, the description is considered distinguishing. However, the description is not distinguishing if there are possible interpretations of the reference objects which do not allow identification of the target (as in Figure 5.1b where interpreting the left square as the reference object would lead to identifying no target at all). Therefore, this interpretation can be seen as a variant of the Independently Unique Reference Object Condition. It shares with it that the identifiability of the reference object is evaluated based on the sub-description referring to it, rather than the entire expression as is done in the Subgraph Matching Condition. On the the other hand, the Total Probability Approach does not require unique identification of the reference object as a precondition for identifying the target.

Reference Object Selection for Probabilistic REG

The intuition behind the subgraph matching approach is certainly convincing in small scenarios. However, expressions generated using this approach may be infelicitous for realistic, cluttered scenes. Figure 5.2 shows an example of a cluttered scene where an RE such as *the circle to the right of the triangle* is arguably discriminating, but not necessarily the most helpful description.

In this case, one would expect an RE to provide assistance for visual search by adding further information which helps identify the reference object, e.g. *the circle to the right of the bottom triangle*.

In an empirical study, Koolen et al. (2015) show that even small amounts of clutter increase the probability of referential overspecification in human subjects. Likewise, Clarke et al. (2013) show in a production experiment with highly cluttered scenes that REs are longer in scenes with more clutter. Moreover, they show that speakers prefer reference objects which are visually salient and in close proximity to the target. These findings indicate that aiding visual search is a relevant factor in REG. In a study using a cooperative building task, Beun and Cremers (2001) found that participants were sensitive to the focus area, often disambiguating objects only within the focus area, and providing longer, more redundant descriptions for objects outside the focus area. Finally, in a study investigating the effect of overspecification on the speed of RR, Arts et al. (2011) show that even in very simple scenarios, superfluous information increases the speed of RR, if it limits the search space. Given this evidence for the relevance of visual search in RR, in this thesis I use an approach to optimality for relational referring expressions which takes into consideration visual search.

The Total Probability Approach seems somewhat intuitive with respect to visual search – if two potential reference objects allow immediate identification of the target, interpreting either of them as the intended reference object would have the same effect, and thus not impede correct identification. However, the allowances made for several potential reference objects run the danger of violating pragmatic expectations. Using the definite article to describe a reference object implies that there exists only one such object. If several objects which fit the description of the reference object are present, this may lead to confusion.

To summarise, there are merits to all three approaches. Clearly, further research is needed in order to evaluate the implications of these approaches and their applicability to human REG in detail. In particular, it may be the case that different principles hold for small scenes vs. cluttered scenes, and for using definite vs. indefinite descriptions for the reference object. As the

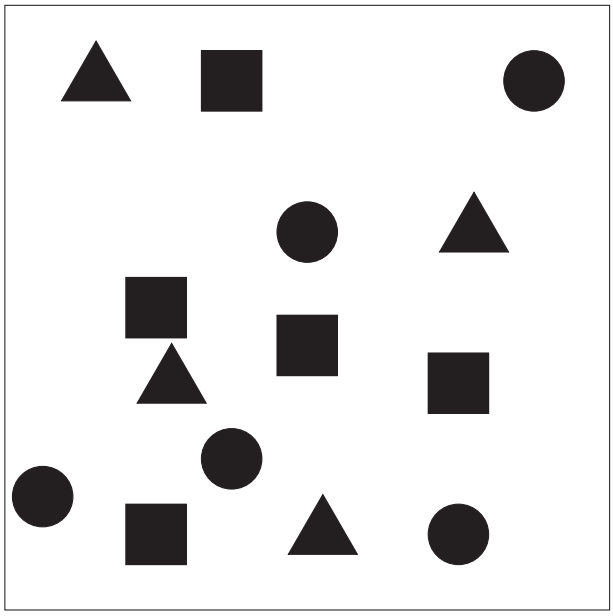


Figure 5.2: Cluttered Scene in which REs may be required to answer WHERE questions as well as WHICH questions.

goal of this thesis is to provide an REG mechanism for situated interaction in realistic scenarios, I will use an implementation of the Independently Unique Reference Object Condition here.

In terms of discriminatory power, this condition can be expressed as the joint probability of identifying both the reference object y and the target object x given the description D: $P(x, y|D)$. Under the assumption that the identification of x is dependent on that of y, and that only the sub-description D' which refers to y is used for identifying y, in this thesis I define the joint probability $P(x, y|D)$ using the following Equation 5.2:

$$P(x, y|D) = P(x|y, D) \cdot P(y|D'). \tag{5.2}$$

In more general terms, in this thesis I evaluate complex REs including n objects in terms of the joint probability, $p(x_1, \ldots, x_n)$, assuming that each reference object must be uniquely identified *before* using it as a reference point for searching for the object described via relation to it.

5.1.3 Further Factors of Reference Object Selection

In the preceding sections, I argued for viewing REG with relations from the perspective of supporting visual search, and therefore taking into consideration the findings of reference object selection research for REG. I then introduced the extension of the PRAGR core concept of Discriminatory Power to handle REs including relations and multiple objects in terms of supporting visual search. By extending the Discriminatory Power definition to handle REs with spatial relations, the issue of reference object selection is integrated into the optimality definition of REG, thus committing to the view that reference object selection in REG cannot be separated from the evaluation of referential descriptions. While it may be reasonable to exclude certain objects from consideration as reference objects based on their inherent properties (e.g., a very small object which is in no way visually salient may not need to be considered as a reference object at all), interactions such as that between the prototypicality of spatial relations and reference object selection (Carlson and Hill, 2009) imply that it is beneficial to be able to directly compare complex descriptions with different reference objects.

However, the individual factors which impact reference object selection remain to be considered, namely locatability of the reference object, search space optimisation, and communication cost. While it is not the goal of this thesis to integrate all aspects of reference object selection, or to investigate their exact interrelations, a number of these aspects follow naturally from the design of PRAGR, or can be easily integrated.

Regarding the factor of locatability of the reference object, I will consider two central sub-factors, namely referentiality and salience, i.e. the ability to provide a description of the object which has a high chance of being understood by the listener, and the degree to which the reference object stands out in the environment.

Referentiality is naturally handled by PRAGR via the concept of Acceptability, as this concept represents the likelihood that a listener would accept the description for the given object. As at this point, PRAGR's handling of Acceptability does not account for acquired knowledge about perceptual or

conceptual differences between speaker and listener. I.e., whether the listener has sufficient knowledge to identify the object based on the description is not considered at this point. Integrating explicit knowledge about the conceptual and perceptual state of the listener into PRAGR and investigating the interrelation between an interlocutor's individual conceptual and perceptual state and grounding processes would be a highly interesting endeavour, but is beyond the scope of this thesis.

As mentioned in Section 3.2, the PRAGR mechanism can be extended to include the salience of objects, giving higher acceptability to descriptions with more salient reference objects. As this topic requires a more in-depth discussion, it will be pursued in detail in Section 5.1.4.

The factor of search space optimisation concerns mainly the prototypicality of the spatial relation used, and the distance between reference object and target object. Both of these factors are taken into account by the projective relation model presented in Section 4.6, which leads to a preference for reference objects which are physically close to the target and in a prototypical angle. As Barclay and Galton (2008) note, the extension of the search space depends on the size of the reference object. The projective relation model covers this to some degree, as the dimension of bounding box angular deviation implies that if a reference object has a large extension on one axis, the Bounding Box Angular Deviation for the relevant relations will be 0 for all target objects which are within the extension of this axis. However, it can be assumed that larger reference objects also imply a larger search space with respect to the acceptable distance of the target object from the reference object. This factor is currently not considered in the projective term model, as the decline of Acceptability based on distance is fixed for the model.

Finally, the factor of communication cost implies a preference for shorter descriptions which is already handled by the PRAGR core mechanism as the weighting parameter α allows determining a preference for more acceptable descriptions. As properties usually do not fit a target perfectly, longer descriptions will automatically lead to a decreasing Acceptability and thus reduce the Appropriateness of a description.

5.1.4 Salience

As was mentioned above, salience plays an important role in reference object selection. In order to guide visual search for the target object, speakers prefer to use reference objects which stick out perceptually or cognitively, i.e., which are more salient than others, in order to help the listener find the intended target object (Clarke et al., 2013).

However, in referential communication, humans also take salience into account in different ways. Humans prefer situationally more salient properties in descriptions (Hermann and Deutsch, 1976). Descriptions for highly salient objects are often shorter than those for non-salient objects (Clarke et al., 2013). Further, listeners resolve ambiguous descriptions by selecting the more salient candidate referents (Clark et al., 1983).

In the following, I will briefly discuss the different kinds of salience and the different ways in which salience can influence REG and RR, followed by a brief overview of existing REG and RR systems which consider salience. I will then show how salience can be integrated into the main PRAGR mechanism. In the following example-based comparison of REs generated by PRAGR with and without consideration of salience, I will show how integrating salience into PRAGR impacts both reference object selection and the length of descriptions. While the impact of salience on RR is also of key interest for an integrated mechanism of reference, this is left for future research.

Kinds of Salience

The concept of salience is hard to delimit, and depending on the perspective, many different definitions have been given. An important distinction which needs to be made is that between bottom-up, visual salience on the one hand, which Itti (2007) defines as "the distinct subjective perceptual quality which makes some items in the world stand out from their neighbors and immediately grab our attention", and top-down, cognitive salience on the other hand, which is goal-driven and depends on the internal state of the viewer (Itti, 2007). However, Itti (2007) emphasises that salience is in all

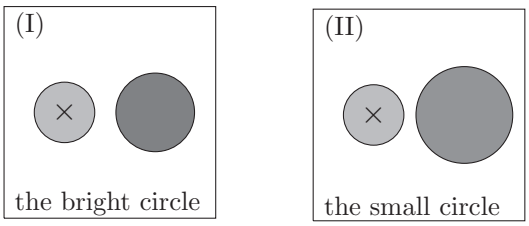


Figure 5.3: Influence of context on property selection in REG (Mast et al., 2016). Situations where (a) brightness, and (b) size is the most salient property.

cases dependent on the viewer, as for example colour blindness means that a person can only perceive salience based on colour to a very limited degree.

Linguistic salience (Kelleher, 2011; Krahmer and Theune, 2002) can be seen as a subtype of cognitive salience. In general, it is assumed that a recently mentioned object will be more salient than one which has not been talked about at all in the current discourse (Krahmer and van Deemter, 2012, pp. 186–188).

An aspect of visual salience which is covered by the main PRAGR mechanism is the preference of humans in RE production for properties which are easy to distinguish. Research on the production of REs by humans shows that contextual differences in the salience of specific object properties impact property choice. For example, Hermann and Laucht (1976) show that when multiple ways of identifying an object are possible, the property with the largest object-distractor contrast, i.e., the most salient property, is chosen. As illustrated in Figure 5.3, humans may prefer to use brightness for an RE in one situation and size in another situation, even if both properties are distinguishing in both situations. The preference is based on the situation specific salience of the different properties.

There is also evidence that speakers adapt the length of REs, depending on the salience of the target object. While Beun and Cremers (2001) found no evidence of an influence of visual salience of a target on the length of REs, they did find that REs for objects that were in the current focus area were less redundant than REs for objects out of the focus area. Clarke et al. (2013) speculate that the failure to find an influence of visual salience on human

production of REs in some studies may be due to the simplified scenarios typically used in such experiments. In a study using highly cluttered scenes, they found that descriptions for non-salient targets were indeed longer than those for highly salient targets.

Salience also serves as an influencing factor for reference object selection in relational descriptions (Barclay and Galton, 2013; Clarke et al., 2013; Gapp, 1996; Talmy, 1983). Barclay and Galton (2013) evaluate the influence of different geometric measures on the selection of reference objects, and conclude that the size of potential reference objects clearly influences selection. Clarke et al. (2013) found that the larger and more visually salient an object, the more likely it is to be used as a reference object in an RE.

From the listener perspective, Kelleher et al. (2005) and Clark et al. (1983) found that human participants use visual salience to resolve ambiguous references, and Strohner et al. (2000) show the same for focus. Frank and Goodman (2012) show that an empirically established measure of object salience leads to successful prediction of the interpretation of REs.

Salience in REG and RR Systems

A number of approaches to REG take salience effects into account by reducing the set of relevant distractors for a description to those which are at least as salient as the intended target (e.g. Jordan, 2000; Krahmer and Theune, 2002; Passonneau, 1996). This allows the generation of shorter, technically underspecified descriptions for salient objects. For example, this would allow *the ball* as a description of an object rather than *the blue ball* despite the presence of other balls in the scene if the ball in question was the most salient ball in the (visual and/or discourse) context.

However, this approach cannot be straightforwardly adapted to REG with graded properties, as it relies on a crisp definition of distinguishing description. Further, it is to be expected that graded category membership interacts with graded salience such that high prototypicality of a description for an object might compensate lack of salience of the object and vice versa. For example, a large green box which is close to the centre of vision may be suf-

ficiently described with the expression *the green box*, even if there are other green boxes in the scene, as long as these are less salient than the intended referent. However, if the most salient green box has a colour which is closer to KHAKI, a less salient, but prototypically green box might be considered to be the intended referent.

The only existing algorithm, to my knowledge, which explicitly integrates salience with graded category membership, is the algorithm for RR proposed by Kelleher (2011). To my knowledge, there is no work on REG which considers salience in reference object selection for REG. Further, to my knowledge there exists no work on salience in the context of REG with graded properties.

The contribution of the present thesis with respect to salience is that of integrating salience into REG with graded properties by integrating a salience measure into the core REG mechanism rather than applying salience effects as a separate step in the generation procedure. In this respect, the approach presented in this thesis most closely resembles the work by Kelleher (2011). However, while Kelleher (2011) focuses on RR, I will focus here on REG, in particular on the influence of visual salience on the length of REs, and on the impact of visual salience on reference object selection.

In the following, I will present the adaptation of the core mechanism of PRAGR for handling salience which will show that salience naturally fits into the PRAGR mechanism. I will then proceed to describe the measure of salience used for the evaluation of the adapted mechanism, before presenting a number of example scenarios and demonstrating how integrating salience impacts the REs PRAGR generates.

Salience in the PRAGR Mechanism

In Section 3.2, I showed how the Discriminatory Power of a description for an object, $P(x|D)$, can be derived from the Acceptability $P(D|x)$ of this description by using Bayes' Rule:

$$P(x|D) = \frac{P(D|x)P(x)}{P(D)}, \tag{5.3}$$

where $P(x)$ is the prior probability of the object, i.e., the probability that, if one randomly chose an object of the given scene, one would pick x. In the simple model presented in Section 3.2, an equal probability of being selected was assumed for each object. However, if we assume that more salient objects will be noticed more easily and therefore have a higher chance of being randomly selected, it is clear that the likelihood of randomly selecting an object depends on its salience. Therefore, in the extension of the model for handling salience, $P(x)$ depends directly on the salience of an object, normalised by the total salience of all objects in the scene:

$$P(x) = \frac{S(x)}{\sum_{x_i \in C} S(x_i)} \tag{5.4}$$

$P(x)$ also feeds into the prior probability of the description $P(D)$:

$$P(D) = \sum_{x_i \in C} P(D|x_i)P(x_i) \tag{5.5}$$

Thus, we can see that this extension follows naturally from the definition of Discriminatory Power given in Section 3.2 and allows easy integration of salience into REG with PRAGR. In the following, I will present an evaluation of REG using this extended mechanism using simple example scenarios.

Calculating Visual Salience

For the evaluation of PRAGR with integrated salience, I used the definition of visual salience provided by Kelleher and van Genabith (2004) for determining visual salience of objects in 3D scenes shown on a 2D screen. Kelleher and van Genabith (2004) determine the visual salience of an object as a function of its centrality in the scene and its size. They use a false colouring technique, rendering the scene normally once, and then a second time using false colours and flat shading such that each object is uniformly coloured in a separate false colour. The image created with false colour rendering can then be used in order to determine which pixels of the viewport belong to each respective object.

In order to determine the visual salience of each object x, Kelleher and van Genabith (2004) calculate the sum of all pixels p of the object weighted by their proximity to the centre of the image:

$$S(x) = \sum_{p \in x} 1 - \frac{\Delta_p}{\Delta_{max} + 1}, \tag{5.6}$$

where Δ_p is the distance between p and the centre of the image, and Δ_{max} is the maximal distance of any pixel in the image to the centre (i.e., the distance between a corner of the image and the centre).

In order to replicate this approach, I created simple scene definitions in tabular form which were then automatically transformed to (a) scene definitions for the POV-Ray 3D rendering software[1], and (b) scene definitions for PRAGR including salience information based on the 3D renderings of the scene. The procedure for generating the required representations was as follows:

1. The tabular scene definition was read in, and the POV-Ray definition for the scene image was created.

2. For each object in each scene, a false colouring version of the image was created in which the object was coloured entirely in black, and surface effects (such as shiny surfaces) were removed.[2]

3. All scenes were rendered using the POV-Ray 3D rendering software.

4. Each false colouring image was processed using a simple script, and the salience of the object was calculated based on Equation 5.6.

5. For each scene, a PRAGR scene definition was output which included the salience information.

[1] http://www.povray.org/

[2] Kelleher and van Genabith (2004) create only one false colouring image for all objects. For technical reasons, I created separate images for each object. The result is the same, as performance issues are not the focus at this point.

Scene	Condition	Description	RO	Acc	DP	App
3	salience	the turquoise book behind the blue mug	RO_{HS}	0.86	0.85	0.85
3	no salience	the turquoise book to the right of the purple mug	RO_{LS}	0.89	0.87	0.87
4	salience	the turquoise book to the right of the purple mug	RO_{LS}	0.89	0.77	0.80
4	no salience	the turquoise book to the right of the purple mug	RO_{LS}	0.89	0.85	0.86
5	salience	the turquoise book to the right of the purple mug	RO_{LS}	0.89	0.80	0.82
5	no salience	the turquoise book to the right of the purple mug	RO_{LS}	0.89	0.87	0.87

Table 5.2: Descriptions generated by PRAGR for the target object in Figure 5.5a where it has high salience and Figure 5.5b where it has low salience. Descriptions with and without consideration of salience are shown.

shorter description for the target object in Figure 5.5a than for the target object in Figure 5.5b. Table 5.1 shows the descriptions generated by PRAGR for the target object in both scenes. Descriptions with and without consideration of salience are shown. As Table 5.1 shows, when considering salience, the description of the target object is indeed shorter for the scene in Figure 5.5a (*the turquoise book*) than for either the same scene without considering salience, or the scene where the target object is less salient. Thus, regarding the length of descriptions, PRAGR with an integrated salience measure produces results that reflect empirical research on RE production and resolution in humans.

Reference Object Selection Figure 5.6 shows scenes with a number of objects (books and mugs) on a table with the respective target objects T. Table 5.2 lists the descriptions generated by PRAGR for the respective target objects, either considering salience or not considering salience. In Figure 5.6a, there are two mugs in the scene, RO_{HS} and RO_{LS}, which are good potential reference objects for this object, as they are in a fairly prototypical relation to it and are both easily identifiable. However, RO_{HS} is more salient due to its more central position closer to the viewer, and RO_{LS} is less salient.

Based on both theoretical considerations and empirical results (Barclay and Galton, 2013; Clarke et al., 2013; Gapp, 1996; Talmy, 1983), an REG system which takes salience into account should clearly prefer the more salient reference object in this situation.

As Table 5.2 shows, when salience is ignored, PRAGR selects the less salient RO_{LS} as the reference object, as it has a slightly more prototypical colour, thus allowing a description with a slightly higher Acceptability value. When considering salience, PRAGR selects the more salient RO_{HS} as the reference object for describing the target object, as the higher salience outweighs the higher Acceptability value. These results show that PRAGR successfully uses salience in order to influence reference object selection.

However, as discussed above, salience is not the only factor that feeds into reference object selection. The factor of *referentiality* listed by Gapp (1995a) indicates that an object which cannot itself be easily described is not a suitable reference object. While there is no direct empirical evidence for this factor in the literature, it is intuitively appealing. As discussed in Section 5.1.2, there are different viewpoints as to whether the reference object needs to be uniquely identifiable independently of the target object (Independently Unique Reference Object Condition), or whether only the configuration of reference object and target object needs to be identifiable. Given the position taken in this thesis favouring the Independently Unique Reference Object Condition, we should expect an REG algorithm to weight the factor of salience against that of referentiality in that sense. A related factor is communication cost – according to Barclay and Galton (2008), a reference object which requires a very lengthy description for successful reference should be dispreferred.

Thus, when a highly salient object allows no short RE which clearly discriminates it from all distractors, we would expect an intelligent REG algorithm to react to this by preferring a less salient, but more easily describable object.

In Figure 5.6b, an additional mug is placed in the scene which is also blue, leading to a situation where the most salient reference object cannot be easily described. In line with the reasoning above, we would now expect

PRAGR to prefer the less salient but more easily describable purple mug as a reference object. As shown by Table 5.2, PRAGR picks up on this change in circumstances and now, as expected, uses the purple mug as a reference object.

Another factor in reference object selection which conflicts with salience is that of search space optimisation. The aspect that large reference objects imply a larger search space is not further evaluated here (although it is covered by PRAGR to some extent). However, empirical findings on reference object selection in RE production show that humans prefer reference objects which are in a prototypical relationship to the target object (Carlson and Hill, 2009). Thus, we would expect that if a highly salient potential reference object were in a non-prototypical relation to the target object, an intelligent REG algorithm would prefer a less salient reference object which is in a more prototypical relationship with the target object.

In Figure 5.6c, the highly salient blue mug is moved such that its projective relation to the target object is less prototypical. We would therefore expect PRAGR to again prefer the less salient purple mug as a reference object, as it is in a more prototypical relation to the target object. As the results in Table 5.2 show, this is exactly what happens: the salience of the blue mug now loses out in favour of the more prototypical relation of the purple mug, yielding again the purple mug as the preferred reference object.

Discussion

The preceding examples show that integrating salience handling into PRAGR has effects similar to the ones found by Clarke et al. (2013) in their study on the production of REs in highly cluttered scenes: REs are shorter for more salient objects, and a preference for salient objects as reference objects emerges. This is all the more intriguing as these assumptions were not explicitly encoded in the way PRAGR handles salience in REG, but rather fall out naturally from modelling the prior probability of an object x via its salience.

While the relationship of this modelling to the preference for more sa-

lient reference objects is fairly straightforward, the fact that PRAGR with integrated salience generates shorter descriptions for highly salient objects is also intriguing, as it is not directly clear how this results from a higher prior probability of the object. However, due to the impact of the prior probability on Discriminatory Power, and the assumed dependence of the identification of the target object on that of the reference object, a reference object which is less salient than the target object is not very likely to increase the Discriminatory Power of a description for the target object, compared to an RE which only uses unary properties to describe the target object. Therefore, if an object is highly salient itself, being surrounded by mostly less salient potential reference objects reduces the likelihood of an RE with relations being used.

Further, the examples above have shown that PRAGR does not simply prefer more salient reference objects, but intelligently balances the factor of salience with other important factors of reference object selection, namely referentiality and search space optimisation, thus displaying the differentiated behaviour predicted by state of the art models for reference object selection. To my knowledge, this kind of sophisticated behaviour regarding reference object selection is unique in the field of REG. While these initial results are promising, empirical studies evaluating both the human-likeness and understandability of REs produced by PRAGR with salience would be required in order to fully evaluate PRAGR's handling of salience and reference object selection.

5.2 Referring Expression Generation as Search

As discussed in Section 3.1, apart from providing a definition of optimality, the generation of REs also requires the use of some search procedure which determines in which order possible descriptions are evaluated.

Reducing computational complexity is usually considered to be a crucial goal of search algorithms. However, a number of issues need to be considered regarding computational complexity.

Firstly, from the perspective of cognitive modelling, computational complexity is only relevant if one proposes a procedural model. As van Deemter (2016, p. 102) points out, most current models are product models in that they attempt to model output similar to that of humans while making no claims regarding the cognitive processes required for producing that output. First steps in this direction have been taken (Gatt et al., 2012), and van Deemter (2016, p. 308) suggests that leveraging the advances of neuroscience may lead to a revival of interest in processing times of REG algorithms. However, from a computational modelling perspective this approach only makes sense if one proposes a procedural model which makes explicit claims about processing times in relation to different stimuli. Moreover, it would imply a vastly different perspective towards computational complexity, as the goal would no longer be to reduce complexity, but to show that an algorithm leads to similar patterns of complexity as displayed by human subjects (van Deemter, 2016, p. 98).

From a practical perspective, van Deemter (2016) notes that the evaluation of complexity may vary, depending on which factors are taken into consideration, e.g., whether one considers complexity with respect to number of objects, number of attributes, length of expression, and/or number of properties. If, for a realistic scenario, a given factor is not expected to exceed a certain number (e.g., one would not expect an RE to consist of more than 100 properties), it may be considered a constant, thus dropping out of the complexity analysis (van Deemter, 2016, p. 98f).

Relatedly, van Deemter et al. (2012a) point out that given the small domains frequently used in REG studies, computational complexity is of-

ten negligible, and full search may be feasible in many cases despite a high theoretical complexity. However, including spatial relations, and assigning non-zero values to all properties – as done in this thesis – puts a higher demand of efficiency on the REG algorithm as the number of properties that may be combined into descriptions increases drastically. Therefore the issue of search needs to be addressed in this thesis.

Due to the inherent directionality of the search problem, this section is concerned exclusively with REG. In RR, the search problem is of lower relevance. While humans do perform visual search when interpreting REs in situated reference, modelling this process is not the focus of this thesis. Therefore, I will consider the search problem only with respect to REG.

In the following, I will discuss the search problem as it occurs in REG, and the implications of the probabilistic optimality measure for the specific search problem faced here. In Section 5.2.1, I will present the basic search framework and how search is handled in the three classic REG algorithms, Full Brevity Algorithm (FB), Greedy Heuristic Algorithm (GH), and Incremental Algorithm (IA). In Section 5.2.2, I will then discuss how the search problem is impacted by extending REG to include spatial relations. The specific problems regarding search that occur when using an optimality definition based on probabilistic properties will be discussed in Section 5.2.3. In Section 5.2.4, I will present the search algorithm used for the implementation of the PRAGR REG component. Finally, in Section 5.2.5, I will perform an evaluation of the computational complexity of REG with the presented algorithm in a number of example scenes.

5.2.1 The Basic Search Framework

The early period of REG research was dominated by three central algorithms: Full Brevity Algorithm, Greedy Heuristic Algorithm and Incremental Algorithm (Dale, 1989, 1992; Dale and Reiter, 1995; Reiter and Dale, 1992). Bohnet and Dale (2005) defined these algorithms as variants of a general search algorithm. Following Russell and Norvig (2002, 65), a search problem can be defined by

1. an initial state to start the search,

2. a successor function which determines possible state transitions for each state,

3. a goal function which determines whether any given state achieves the defined goal, and

4. a path cost function which assigns a cost to the path from the initial state to each reachable state.

The initial state and successor function together determine the state space in the form of a graph with nodes for all reachable states and edges for all possible transitions between states.

In order to solve a search problem, in addition to the problem definition, a search strategy is required: a queuing method which determines in which order possible successor states should be explored. This method may take into consideration values delivered by the path cost function, e.g., by first expanding those states with the lowest costs.

A given search algorithm is called complete if it guarantees that if there is a solution, it will be (eventually) found. It is optimal if it guarantees that the optimal solution is (eventually) found (Russell and Norvig, 2002, p. 74). However, given inherently complex problems, it may be preferable to focus on finding a good solution fast, rather than guaranteeing that the optimal solution is always found (compare Russell and Norvig, 2002, p. 133).

When classic REG is viewed as search, each state consists of (1) a preliminary description – the set of properties of the intended referent that have already been selected, (2) a set of distractors to which the preliminary description applies, and (3) the set of properties that are still available (Bohnet and Dale, 2005). The initial state can be defined as $\langle\{\}, C, P\rangle$, where the context set C is the initial set of distractors, and P is the set of all properties of the intended referent. The goal state can be defined as a valid state which contains an empty set of distractors, indicating a distinguishing description (Bohnet and Dale, 2005).

In the following, I will discuss the three fundamental algorithms of the classic REG, FB, GH and IA, and how they can be described in terms of the search framework.

Full Brevity Algorithm

The Full Brevity Algorithm (Dale, 1989; Dale and Reiter, 1995; Reiter, 1990) focuses radically on the principle of the *minimal distinguishing description*, searching for the distinguishing description with the least number of properties. In terms of the search framework, the FB expands any given node by considering from the set of available properties each which rules out at least one distractor. FB uses a breadth-first queuing method, thus ensuring that the first description which fits the goal condition is the shortest distinguishing description (Bohnet and Dale, 2005). FB is computationally very expensive: given n properties, there are 2^n possible combinations of properties, i.e., the worst case complexity is exponential. Therefore, it is not practically applicable to large problem domains.

Greedy Heuristic Algorithm

The Greedy Heuristic Algorithm (Dale, 1989, 1992) is a more efficient approximation of the FB. This algorithm incrementally builds up a description, utilising the concept of Discriminatory Power discussed in Section 3.1.4: at each point in time, it chooses the property that eliminates the most distractors and adds it to the existing preliminary description, until a distinguishing description has been found. From a search perspective, this algorithm uses a greedy queuing method (hence the name), considering only the property which rules out the most distractors, given the current state. As the greedy algorithm allows no backtracking, it leads to a depth-only search. At each point, the GH has to consider all remaining properties before deciding which one to add, leading to a worst-case complexity of $T(n) = \frac{n(n+1)}{2}$ and thus polynomial complexity ($O(n) = n^2$).

While it has the advantage of efficiency, the GH does not always lead to *minimal distinguishing descriptions*: a property which was added to the

description may be made redundant by subsequently added properties. Dale and Reiter (1995) therefore suggest post-processing the results of GH with a *local brevity* algorithm which takes a distinguishing description, and iteratively forms new distinguishing descriptions from it by either removing a property, replacing a set of properties by a single property, or by replacing a property with a lexically-preferred one. Using this two-fold approach, a *minimal distinguishing description* can be found in polynomial time (Dale and Reiter, 1995).

Incremental Algorithm

The Incremental Algorithm (Dale and Reiter, 1995; Pechmann, 1989) was the most influential REG algorithm of the 1990s, and has had a strong influence on the field until this day (compare Krahmer and van Deemter, 2012; van Deemter, 2016). It is based on empirical evidence that humans produce REs incrementally, preferring the usage of certain properties over others (Pechmann, 1989). The algorithm iterates through all potential attributes in a predetermined order based on assumptions about human preferences. For each attribute, it checks whether the corresponding value of the intended referent eliminates any distractors. If at least one distractor can be eliminated, the property is added to the description and the algorithm proceeds with the next attribute. This process continues, until a distinguishing description has been found. If an object has several possible values for a given attribute – e.g., an object is both a DOG and a CHIHUAHUA, both of which are values of the attribute TYPE – a value is selected with a preference for high discriminatory power and basic level categories.

In terms of search, this algorithm uses an expand method that, for a given state, chooses out of all the properties still available the one which is ranked highest in the predetermined preference order and which rules out at least one distractor. All those properties which do not rule out any distractors are discarded from the list of available properties. This procedure drastically limits the search space, as properties lower in the preference order are not explored independently of the decisions made higher in the preference order.

In the worst case, the algorithm has to check each property exactly once, leading to a worst-case complexity of $O(n)$, making the IA the most efficient of the three algorithms by far.

While the underlying assumption of the IA that humans prefer certain attributes over others is empirically sound (Pechmann, 1989; Viethen and Dale, 2008), the treatment of human preferences in the IA remains simplified. As van Deemter (2016, p. 62f) notes, preferences for certain attributes may be impacted by context, as shown in studies on reference production (Hermann and Laucht, 1976; van Gompel et al., 2014) where subjects preferred attributes which were highly discriminating in the given context. In an empirical evaluation of the IA for human-likeness, van Deemter et al. (2012a) show that its success crucially depends on selecting the correct preference order. With the optimal preference order, the IA outperforms its competitor, the GH. However, with suboptimal preference orders, its performance is significantly worse. This is particularly problematic for domains with many different attributes where a wrong preference order may lead to lengthy, unnatural descriptions (compare van Deemter, 2016, p. 93).

5.2.2 Search Problem with Relations

Extending REG to include spatial relations poses a further layer of challenges for search, some of which have been addressed by extensions of the classic REG algorithms (e.g., Dale and Haddock, 1991; Krahmer et al., 2003; Krahmer and Theune, 2002), while others remain mostly unaddressed.

In the following, I will briefly discuss the challenges to REG with relations which are related to search, namely combinatory explosion, forced incrementality, recursive dependence, and infinite recursion.

Combinatory Explosion

The most fundamental challenge for search in REG with relations is that of combinatory explosion. As Kelleher and Costello (2009) point out, including spatial relations in REG leads to combinatorial explosion already on the property modelling level, as each relation needs to be evaluated once for

each object as a target object with each other object as a potential reference object. Given m objects and n relations this requires modelling $T(m,n) = n \times m \times (m-1)$ acceptability values in total, yielding polynomial complexity for modelling relations alone – $O(m^2)$. If relations need to be evaluated at runtime, partial scene models such as proposed by Kelleher and Costello (2009) may be required. For now, however, I will assume that all relations are modelled for all objects.

Regarding the complexity of REG itself, the situation is even worse. For generating REs containing only unary properties full search has a complexity of 2^n (i.e., there are 2^n possible subsets of properties). When allowing relations, a description for one object may contain property ascriptions to any other object, yielding an upper bound of complexity of $O(2^{n \times m^2})$. This is an upper bound as not all properties are relations, and not all property combinations are valid descriptions (e.g., a valid description cannot contain properties of an object which is not the main target object of a description and not a reference object). However, this upper bound clearly shows that full search for REG with relations is not even viable in small domains. For example, a scene of 10 objects and 10 properties yields $T(n = 10, m = 10) \approx 2^{10 \times 10^2} \approx 10^{301}$.

Forced Incrementality

Given these facts, the popularity of the computationally highly efficient IA and its slightly less efficient sibling GH are not surprising. However, these approaches cause their own specific problems when applied to REG with relations.

Krahmer and Theune (2002) point to the problem of *forced incrementality* which is a major problem for using the IA with relations: in situations where one (less preferred) relation would have been sufficient, the IA may end up concatenating a number of relations which are higher in the preference order but not fully discriminatory, thus yielding very lengthy and awkward descriptions. "The incrementality assumption implies that the first relation will always be realised, even if adding further relations would render it redundant with hindsight. It would seem rather far-fetched to claim psycho-

logical reality for this kind of incrementality." (Krahmer and Theune, 2002). While this problem can be somewhat ameliorated by always assuming relations to be last in a given preference order, it cannot be avoided completely. Further, that solution is far from ideal, given recent evidence that spatial relations are in fact highly frequent in human produced REs (Clarke et al., 2013; Viethen and Dale, 2008). Krahmer and van Deemter (2012, p. 184) reach the conclusion that "relational descriptions [...] do not seem to fit in well with an incremental generation strategy."

Recursive Dependence

A problem with respect to the application of both the GH and the IA to REG with relations is that the recursive nature of REG with relations means that the quality of an RE including a relation crucially depends on the description of the reference object. When greedily selecting a relation due to its high Discriminatory Power, or due to its position early in the preference order, the algorithm does not yet have any information whether the relevant reference object can be successfully described at all. Therefore, applying the GH or the IA to REG with relations risks not finding a distinguishing description at all, even if one exists. This is an inherent problem of all approaches to REG including relations which make local decisions about adding relations to descriptions which do not consider the describability of the reference object. If, on the other hand, one evaluates potential descriptions of the reference object before making a choice, one looses the advantage of greediness. At the least, this problem requires backtracking in case a reference object cannot be described uniquely.

Infinite Recursion

Relatedly, Krahmer et al. (2003) point out the need to prevent infinite recursion which would lead to descriptions such as *the dog in the doghouse that contains a dog that is inside a doghouse...* (Krahmer et al., 2003). Infinite recursion is both a linguistic and computational problem. Linguistically, the resulting descriptions are confusing and not appropriate. Computationally,

infinite recursion may prevent an algorithm from terminating. In their early approach to REG with relations, Dale and Haddock (1991) handle this issue by allowing each property or relation to be used only once. Krahmer et al. (2003) solve the problem more elegantly by using a graph-theoretic approach to REG with relations. They define a scene as a graph such that objects are nodes, and properties are edges. The algorithm then searches for unique subgraphs by recursively expanding existing subgraphs with new edges, and checking for subgraph isomorphism within the larger scene graph. As a graph either contains an edge or not, the problem of infinite recursion does not occur at all in this approach.

5.2.3 The Search Problem in a Probabilistic System

When using a probabilistic optimality criterion for REG as proposed in this thesis, a number of additional problems for search occur. While the basic structure of the search problem still applies, some of the factors which help prune the search space in classic REG do not apply for REG with vague properties.

As discussed above, the traditional REG algorithms define finding a distinguishing description as a stop criterion. While this is by no means required for applying the search paradigm to REG, it does have the clear benefit of providing an unambiguous stop criterion which allows the algorithm to determine whether any given state is a goal state without comparison to other states. The probabilistic approach chosen here does not allow a crisp definition of the distinguishing description and therefore has one clear disadvantage over the classical REG algorithms with respect to search: it is not possible to use the criterion of the distinguishing description as a stop criterion. While it is theoretically possible that the Appropriateness of a description is 1, indicating a perfect description, this is an exceptional case rather than the norm, and most certainly not a requirement in order for a description to be chosen.

One solution for this problem would be to set a threshold value of appropriateness as a stop criterion, thus selecting the first RE which reaches

an appropriateness larger than the threshold value. However, this poses the question of what value should be chosen as a threshold. It is to be expected that the appropriateness of the best description can vary widely both across contexts and domains.

An alternative solution is to perform an exhaustive search over a strongly pruned search tree. If reasonable pruning options are available, pruning the search tree may still yield the best description or a reasonable approximation thereof. The downside of this solution is that in many cases the search will continue long after the best description has been found. Ideally, pruning should make use of the monotonicity assumption in order to prevent further search along paths which are already more expensive than the best solution found so far (Krahmer et al., 2003).

This, however, brings up a further problem of search for REG with vague properties: in classic REG, the path cost function can be defined such that there is a specific positive cost assigned to each step which indicates the cost of using a given property (Krahmer et al., 2003). Assuming that cost is always a positive value, this function fulfils the monotonicity assumption (Krahmer et al., 2003) in that adding a property will always increase the cost of the description.

For probabilistic REG, the path cost function can be defined by the probabilistic appropriateness function described in Section 3.2. As the goal is to maximise Appropriateness, it can be considered a path gain function where the corresponding path cost function would be $1 - appropriateness$. Unfortunately, this definition of the path cost function does not fulfil the monotonicity assumption – the path cost function for a given state S cannot be determined by adding some positive step path cost $c_{S',S}$ to the path cost from the initial state to the prior state $c_{I,S'}$. Thus, adding a property can either increase or decrease the path cost of a description. This poses problems with respect to computational complexity, as a search graph which does not fulfil the monotonicity assumption restricts the possibilities for pruning the search space.

To summarise, the approach described in this thesis can be viewed from a search perspective, although the nature of the search problem proves some-

what different from classic REG and poses a number of additional challenges. In the following, I will present a heuristic search algorithm which overcomes these challenges and is capable of finding an appropriate description in a reasonable amount of time.

5.2.4 A Search Algorithm for Probabilistic REG with Relations

Given the host of complicating factors discussed above, no attempt is made here to present a search algorithm which guarantees finding the most appropriate description given the probabilistic optimality function used in this thesis. Instead, I present a heuristic search algorithm which aims at finding a reasonable approximation in a limited amount of time.

The algorithm used here uses an n-greedy search for the description of individual objects in combination with search space pruning by making use of a tweak which ensures a monotonically increasing path cost function for recursive descriptions involving several objects.

As discussed above, one problem with applying the greedy algorithm to REs including relations is recursive dependence, i.e., the usefulness of a spatial relation to describe a target object relies crucially upon whether the reference object itself can be appropriately described. Or, in probabilistic terms, $P(x, y|D) = P(x|y, D) \cdot P(y|D')$ (see Section 5.1.2). Therefore, in order to fully evaluate the quality of a relational term, one would have to first determine the best description for the reference object, and if that again contains relations, the best description for their reference objects, etc., yielding a depth first search strategy which would require backtracking in the case that no satisfactory description can be found. Further, given the probabilistic path cost function, there would be no clear criterion for determining whether backtracking is indeed necessary, as no crisp definition of distinguishing description can be used (see above).

In the following, I will describe a way of evaluating appropriateness which allows determining whether expanding a given intermediary description by describing its reference objects is worthwhile, thus providing a tool for prun-

ing the search space.

An intermediary description is any description created during search by successively adding properties. We can distinguish two kinds of intermediary descriptions:

1. resolved descriptions (RDs), for which it holds that the target object and each reference object occurring in the intermediary description are described by at least one property contained in the description (see Figure 5.7c), and

2. unresolved descriptions (UDs), for which it holds that the target object or at least one reference object in the intermediary description is not described by any property contained in the description (Figures 5.7a and 5.7b).

One step of resolution is the process in which a unresolved description (UD) is expanded such that one unresolved reference object is described by adding a non-empty set of properties which describe this object. The resulting description may be a resolved description (RD) if all reference objects have been described, or another UD if the initial description contained two unresolved objects, or if a relation with a new reference object was added in the step of resolution. Figure 5.7 illustrates two steps of resolving a description by successively adding a description of an unresolved object.

As described in Section 5.1.2, the probabilistic model proposed in this thesis aims to maximise the joint probability of identifying the target object x and the reference object y. This probability can be determined by the probability of identifying y, given the sub-description which applies to it, and the conditional probability of identifying x, given the full description and that y has been identified: $P(x, y|D) = P(x|y, D) \cdot P(y|D')$. Thus, we can consider $P(x|y, D)$ separately from $P(y|D')$, evaluating only those properties which have x as a target. $P(x|y, D)$ is always equal to or larger than $P(x, y|D)$, as $P(y|D')$ takes a value in the interval $[0, 1]$. Thus, we can use $P(x|y, D)$ as an upper bound of $P(x, y|D)$.

Following this approach, by evaluating UDs which are unresolved for y based on $P(x|y, D)$ we can ensure that no matter how the reference object

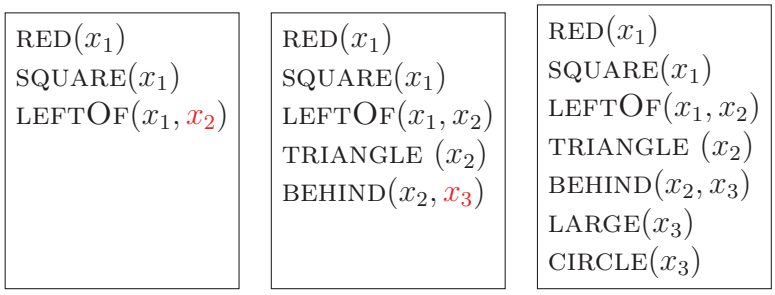


Figure 5.7: Example of stepwise resolution of reference objects: The respective unresolved reference object is marked in red, each subsequent description resolves one unresolved reference object from the prior description. (a) Unresolved description. (b) Unresolved description. (c) Resolved description.

will be described once it is resolved, the appropriateness of the RD will not be larger than that of the UD from which it was obtained by resolution, yielding a path gain function which is monotonically decreasing with respect to resolution steps.

This allows us to prune the search space: any UD with an Appropriateness lower than the best RD found so far can be discarded, as it can impossibly become better by resolution. Only those intermediary descriptions need to be further considered which have a higher appropriateness than the best RD.

This further allows the definition of a general stop criterion for the search algorithm: if each step of resolving an unresolved object involves generating all possible combinations of properties to describe this reference object, we can ensure that after each step of resolution, if the best intermediary description is an RD, this is the best description and the search can be stopped. To give an example, let's assume we start with the empty description and evaluate all possible combinations of properties which describe the target object. If the most appropriate description happens to be an RD (i.e., consists of only unary relations), this is also the most appropriate description overall, as no UDs can possibly become better by performing further resolution steps and thus the preliminary upper bound evaluation of the UDs is sufficient to reach this conclusion. If, on the other hand, the best intermediary descrip-

tion is a UD, the search must continue by performing a step of resolution on the best description. The intermediary description is removed from the list, and all successor states reached by the one step of resolution are added according to their position. The procedure is then repeated with the new best description.

However, as discussed above, generating all possible combinations of properties even for a single target is NP-hard and therefore not desirable. Therefore, we need further measures for pruning the search space with respect to the property combinations that should be evaluated for each individual object.

If we assume that those properties which are individually highly appropriate for the target object are also highly likely to make a relevant contribution to a complex description, we can select the n best individual properties and evaluate only combinations of those. In this case, each step of resolution would yield up to 2^n new descriptions which are inserted into the n-best list before moving to the next step of the algorithm. As n is now a constant, we can limit the complexity of the algorithm by setting a sufficiently small n. This reduction in complexity is achieved at the cost of guaranteeing optimality, as it is conceivable that a property which individually only has a low Discriminatory Power may combine with another property to form the most discriminatory description.

This approach has two weaknesses: Firstly, we are restricted to very small numbers of n, as with higher n the complexity of the algorithm increases exponentially. Secondly, due to the fact that UDs are evaluated based on $P(x|y, D)$ rather than $P(x, y|D)$, the appropriateness of spatial relations is initially overestimated, and this n-best approach may lead to evaluating only descriptions which contain a lot of binary relations and no unary relations. Therefore, in addition to the n best overall properties, the n best unary relations are identified and the two sets combined via set union. This yields an upper bound of complexity of 2^{2n} for one step of resolution.

Given that $P(y|D')$ is independent of $P(x|y, D)$, the task of finding the best description for y can be treated independently of finding the description for x. Thus, we do not need to consider all possible combinations of properties

for x with all possible combinations of properties for y. Instead, when a given UD is to be resolved for the reference object y, and an expansion for y has already been performed in a prior resolution step, this can be directly used for the resolution of the current UD by caching and re-using an ordered list of best descriptions for each target – with one caveat: each resolution step is performed under the consideration of a list of allowed reference objects, as represented in the state. This measure is necessary in order to prevent the occurrence of infinite recursion (see above).

Therefore, a separate list for each set of target and allowed reference objects needs to be cached. if the current UD allows different reference objects for the description of y than prior resolution steps, a new search for resolving y needs to be performed. Otherwise, the already generated list of best descriptions can be re-used.

Each description is integrated into the UD that is being resolved, $P(x, y|D)$ is updated using the formula $P(x, y|D) = P(x|y, D) \cdot P(y|D')$, and the complete description is added to the global sorted list of intermediary descriptions. In the case that a single best description for y, given the allowed reference objects, has already been identified, only a single new description needs to be added. While this procedure does not reduce the theoretical complexity of the algorithm, in practice it yields a major gain, as different combinations of the same relations (with the same reference objects) are being considered and the algorithm need not re-evaluate all possible resolutions of a given reference object for each of these combinations.

To summarise, the search strategy for probabilistic REG using relations is defined as follows:

1. **States:** Each state is defined by the target object, and by an intermediary description consisting of a set of properties that have already been chosen, where each property has a target object and (if it is a relation) a reference object. Further, each state contains a list of objects that are not yet part of the description and are thus available as reference objects for further properties.

2. **Start state:** The start state is defined as an empty set of selected prop-

erties where all objects except for the target object are still available as potential reference objects. As it contains no property describing the target object, the start state is a UD.

3. **Goal function:** The goal function determines that a description must be resolved. If the best description found so far is an RD, the algorithm terminates and returns this description.

4. **Path gain function:** The path gain function is determined by the appropriateness of a given description for the given target object. In the case of RDs, this is based on $P(x_1, \ldots, x_n)$ for all objects x_i which are part of the description. For UDs, the upper bound of appropriateness is calculated based on $P(x_1, \ldots, x_n | u_1, \ldots, u_n, D)$ for all resolved objects x_i and unresolved objects u_i which are part of the description.

5. **Successor function:** Only UDs can be resolved and lead to successor states. A UD is resolved by selecting one of the unresolved objects (i.e., the target object, or any reference object which has not yet been described) and resolving it. If a list of best descriptions for this object and the subsequently allowed reference objects already exists, all descriptions from this existing list are used to resolve the UD. Otherwise, a new list is created. For this purpose, the n individually most appropriate relations (of any arity) and the n individually most appropriate unary relations which describe the object are identified, and a union of these two sets is created. Then, a follower state for each valid subset of this union is created and evaluated. Each object in a description may only be described by one property of each domain, thus preventing descriptions such as *the chihuahua dog* or *the purple blue triangle*. While humans do use descriptions resembling the latter, the implications of such combinations for Acceptability are not entirely clear (an object which is a bad PURPLE and a bad BLUE may be considered a good PURPLE BLUE) and would require further research. Therefore these combinations are not considered here.

6. **Queuing function:** The queue is a list sorted by Appropriateness

in descending order. For RDs, the actual Appropriateness is used, while for UDs the upper bound is used as described by the path gain function. Thus, the algorithm applies a best-first queuing method. For each resolution step, all possible successor states for resolving one unresolved object are created and added to the queue, resulting overall in a breadth-first search.

Algorithm 1 shows the search algorithm used here in pseudo-code where bestDescriptions is the queue as described above. The first element of the list is always the description with the highest appropriateness.

The algorithm may be extended to provide a list of the n best descriptions, in which case the goal function requires that the n best descriptions are RDs. Descriptions are only discarded if their appropriateness is lower than that of the n-best RD that has been found thus far. As long as less than n RD have been found, all UDs are kept for further consideration. Search is continued until the goal function is met, and the n best descriptions are returned.

5.2.5 Evaluation

In order to evaluate the efficiency of the described algorithm in practice, PRAGR was run with 8 scenes containing between 7 and 21 objects. For each scene, all objects in the scene were described by PRAGR, recording the number of resolution steps taken, the number of descriptions considered in total, and the time in ms. All evaluation runs were performed with a maximum number of unary relations and a maximum number of overall relations of $n = m = 5$. Further, descriptions were restricted to allow a maximum of 2 spatial relations for the overall description (i.e., each description contained at most 3 objects - the target object and two reference objects). It needs to be noted that the time measure is only moderately informative, as the evaluation was performed using a laptop running several programs, so interference by other processes may have had an impact. Further, implementation details were not optimised for speed as this is not the focus of this work. The values given for total number of descriptions considered include descriptions taken from re-used best-description lists. While for these descriptions the values of

```
1  bestDescriptions = new SortedList();
   Algorithm searchBestDescription()
2  |  bestDesc = new Description();
3  |  while bestDesc.isUnresolved() do
4  |  |  if nextResolutionStepIsCached(bestDesc) then
5  |  |  |  bestDescriptions.addAll(resolveFromCache(bestDesc));
   |  |  end
6  |  |  else
7  |  |  |  bestDescriptions.addAll(resolveOne(bestDesc));
   |  |  end
8  |  |  cutOffAfter(bestDescriptions,bestResolvedDescription);
9  |  |  bestDesc = bestDescriptions.pop();
   |  end
10 |  return bestDesc;
   Procedure resolveOne(Description desc)
1  |  result = new List();
2  |  subtarget = desc.getNextUnresolvedObject();
3  |  nBestUnaryRelations = getNBestUnaryRelations(subtarget);
4  |  mBest = getMBestRelations(subtarget,allowedROs);
5  |  properties = setUnion(nBestUnaryRelations,mBestRelations);
6  |  bestDescriptionsThisTarget = new SortedList();
7  |  for newDesc in getLegalCombinations(properties) do
8  |  |  newDesc.evaluate();
9  |  |  bestDescriptionsThisTarget.add(newDesc);
10 |  |  if newDesc.isResolved() then
11 |  |  |  cutOffAfter(bestDescriptionsThisTarget, newDesc);
   |  |  end
   |  end
12 |  for newDesc in bestDescriptionsThisTarget do
13 |  |  result.add(combine(bestDesc,newDesc));
   |  end
14 |  return result;
```

Algorithm 1: Search algorithm for generating REs with PRAGR with the main algorithm searchBestDescription and subroutine resolveOne.

Acceptability, Discriminatory Power, and Appropriateness do not need to be calculated afresh, they do need to be considered one-by-one when integrating them into the description as a whole and calculating the combined scores accordingly.

Table 5.3: Mean performance values for $n = 5$ and 2 allowed relations

Total Objects	Resolution Steps	Total Descriptions	Time (ms)
7	8.85	73.70	2.33
8	13.65	108.65	2.60
9	22.17	168.69	3.72
21	55.74	499.33	29.05

Table 5.4: Best (lowest) performance values for $n = 5$ and 2 allowed relations

Total Objects	Resolution Steps	Total Descriptions	Time (ms)
7	1	5	0
8	1	12	0
9	1	12	0
21	1	63	0

Table 5.5: Worst (highest) performance values for $n = 5$ and 2 allowed relations

Total Objects	Resolution Steps	Total Descriptions	Time (ms)
7	29	224	31
8	38	265	16
9	65	346	22
21	138	1,496	78

Table 5.3 shows the mean performance values by total number of objects in the scene, Table 5.4 shows the best (or lowest) performance values achieved by total number of objects in the scene, and Table 5.5 shows the worst (or highest) values reached.

As Table 5.3 shows, the number of resolution steps, number of total descriptions observed, and time taken all increase with increasing size of the scene. However, it needs to be noted that the increase in total descriptions

considered does not increase dramatically with scene size. When roughly doubling the size of the scene (from 9 to 21 objects), the number of total descriptions considered is roughly tripled. Time taken does increase more strongly, a factor which may be due to the fact that for evaluating each description all objects in the scene need to be considered – a linear factor which may play a crucial role for practical run-time.

As Table 5.4 shows, at all scene sizes the best-case scenario requires only one step of resolution, thus identifying a resolved description as the best description immediately, in less than $1ms$. The difference in the number of descriptions considered in this one step of resolution is caused by the set of individually appropriate properties that are determined prior to the combination. If the set of best unary relations and the set of best relations overall coincide, only n properties are considered. If most of these belong to the same attribute (e.g., colour), only a very small number of combinations will be considered due to the constraint that only properties belonging to different attributes may be combined (with the exception of relations, as e.g. two different projective relations may be combined). If, on the other hand, most best relations overall are not unary relations, up to $n + m$ individual properties are considered. Further, if most properties belong to different attributes or are binary relations, a much larger number of combinations is possible.

As Table 5.5 shows, the worst case for the test cases considered is roughly 2 to 3 times the mean, and thus does not cause any serious issues. However, it must be noted that the number of test cases is very small, and a more thorough investigation would be necessary in order to evaluate whether worst-case performance might cause problems in applications.

Overall, the evaluations with the strongly restricted n and m values, and allowed relations give cause for optimism and warrant further investigation with less restrictions.

Table 5.6 shows the mean performance metrics when increasing n and m to $n = m = 10$ while still allowing a maximum of 2 spatial relations per description. Table 5.7 shows the mean performance metrics when leaving $n = m = 5$ while allowing up to 5 spatial relations per description. Table 5.8

Table 5.6: Mean performance values for $n = 10$ and 2 allowed relations

Total Objects	Resolution Steps	Total Descriptions	Time (ms)
7	16.35	437.70	11.30
8	31.31	853.50	13.33
9	70.61	1,295.69	34.22
21	210.62	4,232.17	198.62

Table 5.7: Mean performance values for $n = 5$ and 5 allowed relations

Total Objects	Resolution Steps	Total Descriptions	Time (ms)
7	10.68	115.33	4.50
8	25.12	306.50	4.96
9	60.28	868.28	11.22
21	359.38	10,730.90	189.88

shows the mean performance metrics when increasing n and m to $n = m = 10$ and allowing up to 5 spatial relations per description. As the tables show, increasing n or the number of allowed relations individually strongly increases the average number of steps and the total number of descriptions considered. However, reasonable efficiency is still retained. However, increasing both n and the number of relations allowed together, leads to unacceptable performance with up to $3,858,096$ descriptions considered (see Table 5.9) and processing times of several seconds.

5.2.6 Discussion

In this section, I have presented a definition of the search problem of REG which is compatible with both relational properties and a concept of Discriminatory Power based on vague properties.

The problem of forced incrementality does not occur with the algorithm presented as the evaluation measure always considers the description as a whole and does not incrementally add properties which may later become redundant.

Further, the upper bound evaluation for UDs provides a useful tool for handling recursive dependence, thus preventing the kinds of problems the

Table 5.8: Mean performance values for $n = 10$ and 5 allowed relations

Total Objects	Resolution Steps	Total Descriptions	Time (ms)
7	17.18	918.55	21.02
8	63.02	4,521.31	68.40
9	312.78	17,570.50	224.19
21	3,979.64	1,042,446.48	8,213.55

Table 5.9: Worst (highest) performance values for $n = 10$ and 5 allowed relations

Total Objects	Resolution Steps	Total Descriptions	Time (ms)
7	65	2,621	110
8	225	12,757	160
9	2,400	112,542	826
21	16,457	3,858,096	14,386

GH runs into when confronted with several potential reference objects.

By keeping track of available reference objects in the state representation, the problem of infinite recursion is avoided.

With the algorithm presented here, the rampant combinatory explosion caused by allowing binary relations in REG can be kept at bay. The experimental performance analysis has shown that the complexity of the resulting search algorithm is still higher than desirable. However, restricting either the number of relations allowed in a description, or the number of individually appropriate properties being considered for inclusion in the description can keep complexity at reasonable levels for relatively complex scenes. The first restriction (number of allowed relations) limits the kind of expressions that can be generated. However, it is questionable in any case that a human would use a description containing long chains of reference objects. The second restriction (number of individually appropriate properties considered in full description) concerns the danger of not finding an optimal solution due to eliminating a property which, while not being very appropriate in and of itself, would increase the appropriateness of an entire description. Based on the descriptions generated in the empirical evaluation discussed in Chapter 6, this is not a particularly pressing issue. However, it should be kept in

mind.

To sum up, while certainly not perfect, this algorithm demonstrates that the issues which binary relations and vague property representations cause for search in REG can be tackled, and the PRAGR mechanism described here is not only theoretically interesting, but applicable in practice. However, further research on heuristic approaches that could further improve efficiency is warranted, and some options are discussed in Section 8.3.4.

5.3 Summary

In this chapter, I have discussed the core challenges faced when integrating spatial relations into an REG algorithm and demonstrated how PRAGR can tackle the most relevant of these issues in keeping with the approach to vagueness taken here.

In particular, I have integrated the identifiability of the reference object into an overall evaluation of Discriminatory Power based on the assumption that REG has the goal of supporting visual search. Further, I have integrated salience into the REG mechanism, and presented a search algorithm which finds appropriate descriptions in a reasonable amount of time.

In conclusion, the contribution of this chapter was to demonstrate that the PRAGR mechanism is capable of handling a range of relevant challenges in REG in an integrated fashion and can overcome some of the shortcomings of earlier REG algorithms in the areas of REG with spatial relations, handling salience, and search for REG.

With this chapter, the presentation of the PRAGR mechanism itself has been completed, and I will proceed to present a number of empirical evaluation experiments which serve to evaluate the performance of PRAGR in realistic scenarios before discussing the application of PRAGR in referential dialogues.

Chapter 6

Evaluation

In this chapter, I will present three empirical studies evaluating the performance of PRAGR in scenarios involving robot-robot and human-robot interaction. For the evaluation studies, PRAGR makes use of several different property models, including spatial relations in order to demonstrate the ability of PRAGR to handle the interaction of a variety of property models. In order to evaluate the claim that vague property models are superior to crisp ones, each evaluation study is performed with two conditions, using either vague property models or their crisp counterparts, as explained in Section 4.8.

In Section 6.1, I will briefly discuss prior evaluation challenges for REG and metrics used for evaluating REG before continuing to describe the evaluation studies in detail in the following sections. Section 6.2 will cover the evaluation of PRAGR in robot-robot interaction under conditions of perceptual deviation in order to evaluate the ability of PRAGR to enable referential grounding dialogues, and in order to evaluate the usefulness of vague properties for overcoming perceptual deviation and supporting grounding dialogues. In Section 6.3, I will present an evaluation of reference resolution (RR) where PRAGR interprets human produced utterances. An evaluation study testing human performance resolving REs generated by PRAGR will be described in Section 6.4.

6.1 REG Challenges and Evaluation Procedures

In the past, a number of REG Shared-Task Evaluation Competitions (STECs) have been conducted in which several systems were evaluated against a joint dataset in order to enable comparison. With the ASGRE challenge in 2007, REG was the first area of NLG to be tackled with an STEC (van Deemter, 2016, p. 107).

Between 2007 and 2009, three competitions were conducted using the TUNA dataset (Gatt et al., 2009) which consists of pairings between visual scenes with one or two target objects and 6 distractors on the one hand, and human produced REs for the target objects on the other hand. For the purpose of REG, objects are represented as attribute value pairs. The dataset consists of two subsets, one with images of furniture, and another one with portrait photographs of humans. Evaluation in the TUNA-REG'09 challenge, the third challenge using the TUNA corpus, was performed using a mix of automatic evaluation metrics, human judgment, and task success measures (Gatt et al., 2009).

The GREC challenge (Belz et al., 2008) evaluated generation of REs for coherent texts using 2000 texts from introductory Wikipedia articles with manually conducted annotations.

In contrast to those studies which are restricted to pure REG, the GIVE and GRUVE challenges provided integrated NLG challenges in which automatically generated instructions were evaluated for task success in navigation tasks (Koller et al., 2010; Striegnitz et al., 2011).

However, except for the GRUVE challenge, all these joint evaluation challenges were based on predetermined crisp properties, offering no adequate testing ground for evaluating PRAGR. Therefore, all stimuli and human data used in the following evaluations were created or collected specifically for this purpose.

In Section 3.1.1, I discussed the general approaches that can be taken towards optimality in reference: human-likeness vs. task success. These are reflected in different kinds of evaluation measures. Gatt et al. (2009)

distinguish three types of evaluation measures based on what aspects of performance are measured and how: automatic intrinsic, human intrinsic, and extrinsic. Intrinsic measures evaluate the RE in and of itself, while extrinsic measures evaluate the RE in terms of success at a given task.

Automatic intrinsic evaluation concerns only human-likeness, using automatic test scores such as string edit distance (SE), the BLEU-x score which performs string comparison based on n-grams, or NIST which extends the BLEU-x score to weight n-grams according to their frequency.

Human intrinsic evaluation may cover different aspects of quality. It is conducted by asking human judges to evaluate an expression with respect to different properties. For example, Gatt et al. (2009) asked human participants to judge how clear and how fluent a description was using a sliding scale.

Extrinsic measures concern any performance measures which evaluate the quality of an expression based on the performance of a human in a given task. In REG, this is usually task success in RR – whether or not the listener is able to correctly identify the target from a given description. Gatt et al. (2009) also measure the speed of identification.

As discussed in Section 3.1.1, the core criterion for optimality used in developing the presented mechanism was that of task success, therefore the evaluation will also focus on task success. As one core goal of this thesis is to present a unified mechanism for generation and resolution of REs as a basis for enabling grounding dialogues, evaluation covers both these areas. As the main contribution of this thesis is a mechanism for handling vague properties, all evaluation studies will compare PRAGR using vague properties with a version using corresponding crisp properties.

6.2 Evaluation in Robot-Robot Interaction

The first evaluation concerns the general capability of PRAGR to generate and understand REs for objects in a visual scene, and the potential of the mechanism to handle situations of perceptual deviation, focusing on the communicative potential of crisp versus vague categories in reference. For this

Figure 6.1: Experimental setup with two NAO robots jointly observing a scene (Mast et al., 2016).

end, an experimental setup was used where two robots view the same scene from slightly different positions and play a language game (Steels et al., 2005). In this language game, one robot describes an object and another attempts to resolve the intended referent of the RE. Figure 6.1 shows the experimental setup involving two Nao robots that jointly view a scene of geometric objects from different perspectives. Each robot records the scene using an integrated camera (bottom head camera with 640×480 pixel resolution). The evaluation of the images collected this way was performed offline on an external computer in order to allow replicability of the results.

Overall, the implementation used for this experimental setup consists of the following components:

- two Nao humanoid robots with an internal video camera,
- an object segmentation component,
- several property models,
- two separate instances of the PRAGR reference handler (speaker and listener), and
- an experiment handler.

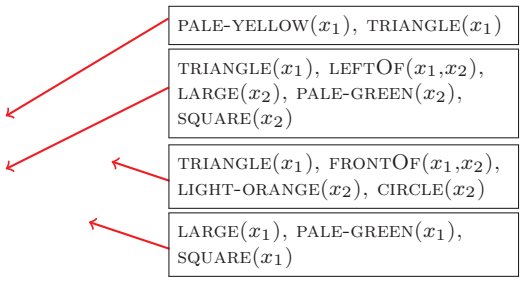
In the following, I will describe the design of the stimuli used, the overall experimental procedure, the object segmentation and feature extraction

methodology, and the property models used. I will then proceed to present the results and discuss the implications of the experiment.

6.2.1 Stimuli

The experimental scenes were created by automatically generating 100 shape objects (circles, triangles, and squares) of different colours. The colours of the 100 objects were randomly generated, with an emphasis on colours with lower saturation, making the task harder. The area of the largest shapes was approximately 2.5 times the area of the smallest shapes. From these 100 shape objects, 48 images were created. The number of objects in each image was set to a random number between 4 and 10, and each object selected at random from the set of available objects, yielding a list of scene definitions consisting of 4-10 object identifiers. Objects were not allowed to appear more than once in one image. The configuration of the objects in the scene was performed spontaneously by the experimenter prior to capturing the image: throughout the image collection phase, the two robots were positioned at a slight angle, viewing approximately the same area on the ground. The experimenter selected the objects which corresponded to the identifiers given in the scene definition and placed all objects on the area visible to both robots. Then, each robot recorded a picture of the scene. Figure 6.1 shows the positioning of the robots and the scene. This combination of automatic random scene creation with human intervention (positioning of the objects) was chosen in order to minimise the impact of human bias on the stimuli while ensuring practicality of the procedure. All scenes together contained a total of 327 objects, yielding an average of 6.81 objects per image.

For each scene, a pair of images was recorded – the scene as seen by the robot Alex, and the scene as seen by the robot Amy. In order to establish the correspondence between objects in the two different views of the scene for the purpose of evaluation, object segmentation was performed in advance of experimentation, yielding fixed IDs for all objects. To obtain ground truth for automatic evaluation, all image pairs showing the same scene were then manually annotated for correspondence between object IDs in the different



(a) (b)

Figure 6.2: Example of scene as seen by Alex, with visualisation of object segmentation and referring expressions generated for objects in the scene. (a) Alex' segmentation of (b). (b) Camera image of robot and some generated descriptions (images adapted from Mast et al., 2016).

images, e.g., Alex' object 1 may be identical with Amy's object 4. This way, it was then possible to automatically evaluate after each interaction whether the intended target object had been identified correctly.

6.2.2 Object Segmentation and Feature Extraction[1]

Due to the simple nature of the stimuli, a lightweight object segmentation component was sufficient for the purpose of this experiment. The scenes contain no partially occluded objects or objects with complex texture which cannot be easily separated from the background. Further, all objects are unicoloured.

Object segmentation was realised by searching for contours in the image, considering local contrast in either hue or lightness. The implementation is based on the OpenCV library and applies the Canny edge detector for contour finding (Canny, 1986). Object contours get distorted by projection on the camera, but can be approximately corrected by applying camera calibration and perspective transformation according to the estimated head orientation

[1]Object segmentation including correction of distortion and extraction of perceptual features was contributed by Diedrich Wolter and is not part of the contribution of this thesis.

with respect to the ground plane. Figure 6.2a shows an example of the segmented image with object IDs and the most likely shape property.

As discussed in Section 4.6, a projected relative reference frame was assumed for the experiments, basing the evaluation of projective terms on the viewing direction of the speaker, projected onto the reference object, as shown in Figure 6.3a. Thus, in order to model projective terms, the listener needs to adapt their perspective to that of the speaker. This can be roughly approximated by estimating the difference in the viewing angle between speaker and listener and rotating the entire scene by this estimated difference. Figure 6.3b illustrates the adaptation of the perspective by the listener. Thus, for the pictures taken by Amy, a perspective adaptation was performed by rotating the segmented image according to the estimated perspective difference used in the different conditions (between 20° and 50°, see below), resulting in a separate segmented image for each rotation.

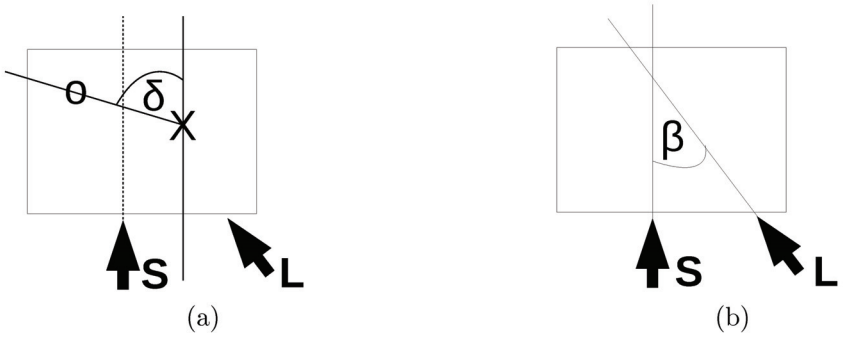


Figure 6.3: Reference frame and perspective adaptation used for evaluation. (a) Frame of reference within image as determined by orientation of speaker S. (b) Perspective adaptation by listener L is based on β: the difference between L's and S's viewing direction (Mast et al., 2016).

Based on the contour retrieved from segmentation, for each object the minimal distances to the shape prototypes (SQUARE, TRIANGLE, CIRCLE) were calculated using the shape model described in Section 4.5. The colour of each object in terms of hue, saturation, value (HSV) was extracted by taking the mean value of a $7px \times 7px$ square centred on the centre of mass of the object. The value was later transformed to Hue, Saturation and Lightness

(HSL). Further, the contour was used for extracting the size of the object in pixels, and the bounding box which was used for modelling projective relations.

The extracted values of each scene were stored in a text file for processing with further property models and for generating descriptions using PRAGR.

6.2.3 Property Models

The following property models were used (the number of the model in Table 4.1, page 95 is given in brackets):

- full colour model with 37 colour terms (model 6, see also Section 4.4)
- contour based shape model with the concepts SQUARE, TRIANGLE, CIRCLE (model 8, see also Section 4.5)
- size model with concepts LARGE and SMALL, using local prototypes (model 4, see also Section 4.3)
- projective relation model with horizontal relations (model 9, see also Section 4.6)

For all models, in addition to the vague model, a corresponding crisp model was created by performing Voronoi tessellation based on the acceptability values, as described in Section 4.8.

6.2.4 Experimental Setup

For each experiment, two instantiations of PRAGR and the corresponding property models were run in parallel, connected by the experiment handling component (henceforth termed *experiment handler*). Throughout the experiment, Alex took the role of speaker, while Amy took the role of listener. The experiment handler presented the respective scene to each participant, who, after some minor preprocessing (e.g., transforming HSV values to HSL values) each called the property models in order to receive acceptability values of all properties. Then, for each object in the scene, Alex generated

a description using PRAGR, selecting the appropriate set of attribute-value pairs. The description was transferred to Amy by the experiment handler, via a semantic representation, bypassing surface realisation and parsing. Then Amy determined the most probable target object using her separate instance of PRAGR and returned this to the experiment handler which then used the manually created correspondence table to check whether the correct target was identified. For the descriptions, a maximum of one spatial relation was allowed, while the number of other properties was not further limited.

The experiment was run in several configurations. Firstly, use of vague vs. crisp property models was varied: pairs of speaker and listener using only vague property models (Vague-Vague condition) were compared to pairs using only crisp property models (Crisp-Crisp condition). In order to tease apart whether vagueness on the side of the speaker or the listener had more of an impact, additional experiments were run with asymmetric configurations, i.e., vague speaker and crisp listener (Vague-Crisp condition), and vice versa (Crisp-Vague condition).

Secondly, the perspective adjustment angle estimated by Amy, the listener, was varied. The real difference in viewing angles between the robots was roughly 30°. Estimation angles between 20° and 50° were tested in order to simulate more or less accurate estimation.

Thirdly, PRAGR's model parameter α was varied using values between 0.25 and 0.35, indicating a preference for descriptions which are more discriminatory (lower α) or more acceptable for the object per se (higher α).

Finally, the configuration of models was varied, using only a single property model (colour, volume, or shape), a combination of those three, or all three in addition to projective relations.

6.2.5 Results

The complete results of the evaluation are shown in Table 6.1. Figure 6.4 summarises the results using all available property models and a rotation adjustment of 35° for the different possible combinations of crisp vs. vague models in speaker and listener across different values for α. The vague prop-

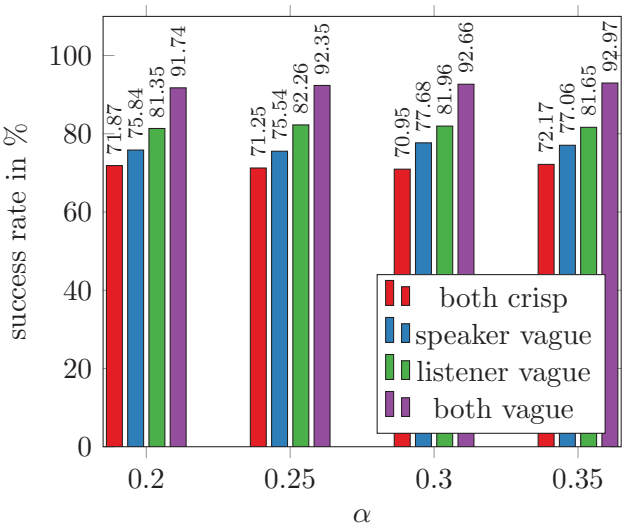


Figure 6.4: Success rates of correct object identification for robot-robot communication across different values of α using vague vs. crisp property models. Rotation adjustment: 35°.

erties consistently produce better results than the crisp models, irrespective of α. Overall, as shown in Table 6.1, when considering only conditions using all property models, the results of the Vague-Vague condition range from 84.4% to 93.27%, with $\alpha = 0.2$ and $rotationAdj = 40°$ reaching the highest score (305 out of 327 objects correctly identified by the listener). Scores for the Crisp-Crisp condition range from 68.2% to 72.17%, with $\alpha = 0.35$ and $rotationAdj = 35°$ performing best (236 out of 327 object correctly identified). Using vague speakers with crisp listeners (Vague-Crisp condition) improved performance compared to the Crisp-Crisp condition, scoring up to 75.84% correct identifications. Crisp speakers with vague listeners (Crisp-Vague condition) showed even better results, scoring up to 82.26%. As Figure 6.4 shows, the variation of α has only a minimal effect on the results.

Figure 6.5 summarises the results for including different individual property models or combinations of property models. These results reveal how informative the different domains are, and how much the system benefits from being able to choose and combine domains. For each property alone, results are fairly low irrespective of whether crisp or vague models are used, with the exception of the COLOUR attribute which yields success rates of over

Table 6.1: Evaluation results of referential robot-robot communication in different settings.

α	speaker	listener	rotation [°]	Volume	Colour	Shape	Spatial	correct	incorrect	success [%]
0.20	crisp	crisp	35			×		100	227	30.58
0.20	crisp	crisp	35		×			170	157	51.99
0.20	crisp	crisp	35	×				78	249	23.85
0.20	crisp	crisp	35	×	×	×		215	112	65.75
0.20	crisp	crisp	20	×	×	×	×	223	104	68.2
0.20	crisp	crisp	25	×	×	×	×	225	102	68.81
0.20	crisp	crisp	30	×	×	×	×	229	98	70.03
0.20	crisp	crisp	35	×	×	×	×	235	92	71.87
0.20	crisp	crisp	40	×	×	×	×	232	95	70.95
0.20	crisp	crisp	45	×	×	×	×	229	98	70.03
0.20	crisp	crisp	50	×	×	×	×	229	98	70.03
0.25	crisp	crisp	35	×	×	×	×	233	94	71.25
0.3	crisp	crisp	35	×	×	×	×	232	95	70.95
0.35	crisp	crisp	35	×	×	×	×	236	91	72.17
0.20	vague	crisp	35			×		100	227	30.58
0.20	vague	crisp	35		×			126	201	38.53
0.20	vague	crisp	35	×				80	247	24.46
0.20	vague	crisp	35	×	×	×		190	137	58.1
0.20	vague	crisp	20	×	×	×	×	242	85	74.01
0.20	vague	crisp	25	×	×	×	×	241	86	73.7
0.20	vague	crisp	30	×	×	×	×	244	83	74.62
0.20	vague	crisp	35	×	×	×	×	248	79	75.84
0.20	vague	crisp	40	×	×	×	×	245	82	74.92
0.20	vague	crisp	45	×	×	×	×	245	82	74.92
0.20	vague	crisp	50	×	×	×	×	240	87	73.39
0.25	vague	crisp	35	×	×	×	×	247	80	75.54
0.3	vague	crisp	35	×	×	×	×	254	73	77.68
0.35	vague	crisp	35	×	×	×	×	252	75	77.06
0.20	crisp	vague	35			×		100	227	30.58
0.20	crisp	vague	35		×			191	136	58.41
0.20	crisp	vague	35	×				94	233	28.75
0.20	crisp	vague	35	×	×	×		251	76	76.76
0.20	crisp	vague	20	×	×	×	×	261	66	79.82
0.20	crisp	vague	25	×	×	×	×	266	61	81.35
0.20	crisp	vague	30	×	×	×	×	269	58	82.26
0.20	crisp	vague	35	×	×	×	×	266	61	81.35
0.20	crisp	vague	40	×	×	×	×	264	63	80.73
0.20	crisp	vague	45	×	×	×	×	268	59	81.96
0.20	crisp	vague	50	×	×	×	×	265	62	81.04
0.25	crisp	vague	35	×	×	×	×	269	58	82.26
0.3	crisp	vague	35	×	×	×	×	268	59	81.96
0.35	crisp	vague	35	×	×	×	×	267	60	81.65
0.20	vague	vague	35			×		100	227	30.58
0.20	vague	vague	35		×			266	61	81.35
0.20	vague	vague	35	×				111	216	33.94
0.20	vague	vague	35	×	×	×		292	35	89.3
0.20	vague	vague	20	×	×	×	×	276	51	84.4
0.20	vague	vague	25	×	×	×	×	284	43	86.85
0.20	vague	vague	30	×	×	×	×	294	33	89.91
0.20	vague	vague	35	×	×	×	×	300	27	91.74
0.20	vague	vague	40	×	×	×	×	305	22	93.27
0.20	vague	vague	45	×	×	×	×	305	22	93.27
0.20	vague	vague	50	×	×	×	×	300	27	91.74
0.25	vague	vague	35	×	×	×	×	302	25	92.35
0.3	vague	vague	35	×	×	×	×	303	24	92.66
0.35	vague	vague	35	×	×	×	×	304	23	92.97

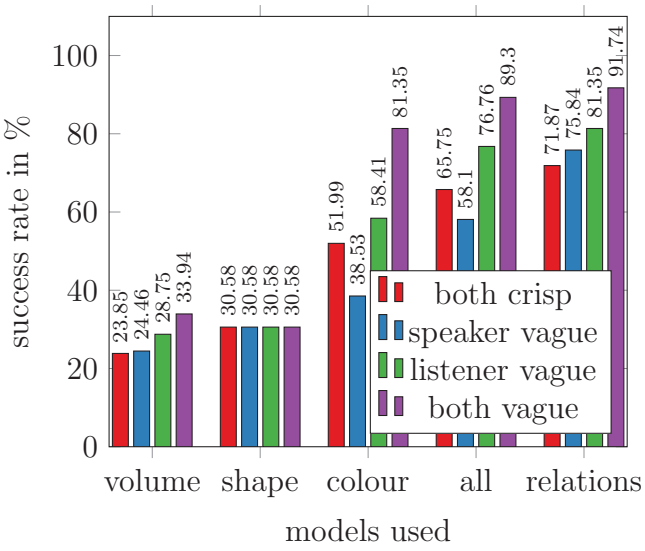


Figure 6.5: Success rates of correct object identification for robot-robot communication using vague vs. crisp property models, depending on the attributes covered. Rotation adjustment: 35°, $\alpha = 0.2$.

50% for most configurations. Notably, there is no difference between the crisp and the vague models for SHAPE only (30.58% for all configurations), most probably due to the fact that the shape model was adapted to reflect the relatively crisp nature of human perception of the shapes in question (see Section 4.5.2). For SIZE, there is a substantial difference based on crispness of models (Vague-Vague 33.94%, Crisp-Crisp 23.85%), and for colour the difference is even more pronounced (Vague-Vague 81.35%, Crisp-Crisp 38.53%). Interestingly, the Vague-Crisp condition yields worse results for the COLOUR than the Crisp-Crisp condition, indicating that using vagueness in speaking is not beneficial for colour if the listener is not vague. In contrast, the Crisp-Vague condition slightly improves over the Crisp-Crisp condition.

Combining all three property models (without spatial relations) yields better results than the best single model in all conditions, with the increase in the Crisp-Crisp condition being higher than in the Vague-Vague condition, and the increase in the Vague-Crisp condition being highest. (Vague-Vague 89.3%, increase of ca. 8 percentage points, Crisp-Crisp 65.75%, increase of ca. 14 percentage points. Vague-Crisp 58.1%, increase of ca. 20 percentage

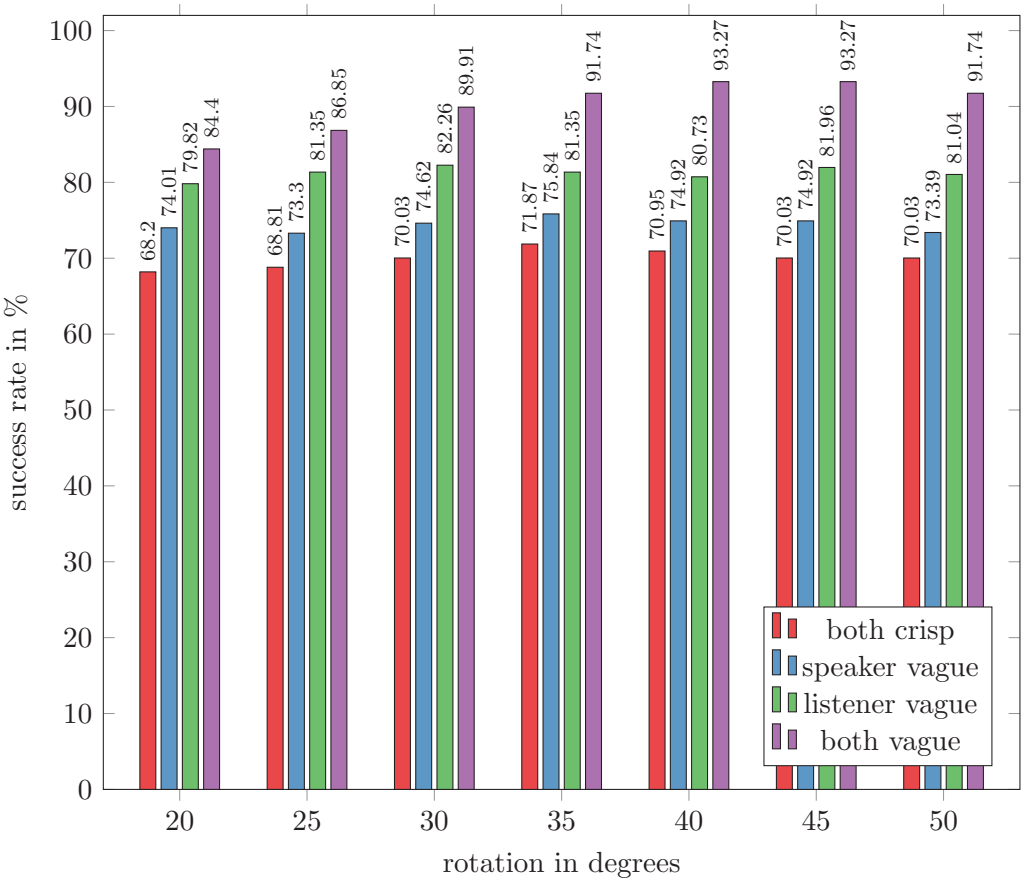


Figure 6.6: Success rates of correct object identification for robot-robot communication across different estimated rotation angles using vague vs. crisp property models. $\alpha = 0.2$.

points). Adding the option of using spatial relations yields slight benefits for all conditions. Again, the Vague-Crisp condition benefits the most, mostly recovering from the very bad performance of the colour model.

Figure 6.6 summarises the results for different rotation adjustments using $\alpha = 0.2$. The Vague-Vague condition outperforms all other conditions across all rotation adjustments, with the Crisp-Vague condition next, followed by the Vague-Crisp condition and finally the Crisp-Crisp condition which fares worst. Superficially, changing the degree of rotation has a stronger impact on the Vague-Vague condition than on the other conditions. In the Vague-Vague condition, correctness drops from 93.27% for the best performing rotation ad-

justment (40°,45°) to 84.4% in the worst case (20°), corresponding to a drop of ca. 9 percentage points. In the Crisp-Crisp condition, correctness drops from 71.87% for the best performing adjustment (35°) to 68.2% in the worst case (20°), corresponding to a drop of ca. 3.5 percentage points. However, closer inspection showed that in the case of the vague models, a high number of descriptions (typically > 200 out of 327) contained a spatial relation, making the system more vulnerable to incorrect estimation of perspective deviation. In the crisp condition, typically less than 100 descriptions contained a spatial relation, indicating that the spatial relations were not considered informative by the speaker. Thus, the lower vulnerability to perspective deviation may be attributed to the lack of spatial relations used.

6.2.6 Discussion

The results show that for robot-robot communication, vague property models improve communicative success considerably, and thus support the main hypothesis of this thesis, that situated referential communication in the face of perceptual deviation benefits from using vague property models. Moreover, the success rates of up to 93.27% for a difficult setting (many similar shapes, mostly low saturation colours, no unique landmarks) demonstrate that the generalisation of the concept of discriminatory power to graded properties, as realised by PRAGR, can strongly improve referential success over approaches based on crisp categories.

The improvement gained by combining the different models within one mechanism shows that PRAGR is capable of making reasonable decisions between different available conceptual domains, leaving out confusing information while adding informative properties. However, this improvement was achieved with both crisp and vague models, thus the central improvement of adding vagueness does not seem to be the better handling of combinations, but rather the better base performance of the models, especially the colour model.

6.3 Evaluating the Understanding of Human-Produced Descriptions

In order to evaluate PRAGR's ability to interpret REs produced by humans, I collected a corpus of human descriptions of objects using the same scenes that were used for the robot-robot evaluation. I evaluated PRAGR's interpretation of the descriptions under the same condition of perceptual deviation as in the robot-robot scenario.

Human subjects saw the scene as seen by Alex and were asked to provide a description of an object. The system took the role of the listener, receiving pairs of scene photos as seen by Amy and human-produced descriptions. The system parsed human produced descriptions using a simple parsing mechanism and generated a list of most likely targets.

Overall, the implementation used for this experimental setup consists of the following components:

- the description collection software,
- an object segmentation component,
- a simple text parser,
- several property models,
- one instance of the PRAGR reference handler (listener), and
- an experiment handler.

In the following, I will describe the procedure for gathering data, the property models used, the parsing process, and the interpretation and evaluation procedure. I will then proceed to present the results and discuss the implications of the experiment.

6.3.1 Data Collection

Overall, 13 people aged 24 to 49 (mean: 32) participated in the study (6 female, 7 male), predominantly students or university graduates. All par-

rötlichbraunes Dreieck *(reddish brown triangle)*
das lilagraue dreieck rechts neben dem grünen dreieck
(the purple grey triangle to the right next to the green triangle)
das Dreieck ganz rechts außen *(the triangle right on the outer right)*
Dreieck rechts neben dem grünen Dreieck und oberhalb von dem hellroten Kreis
(triangle to the right next to the green triangle and above the lightred circle)
das rötlich-graue Dreieck rechts, oberhalb vom roten Kreis
(the reddish-grey triangle on the right, above the red circle)
mittelgroßes helles mattes graues lilanes dreieck
(medium-sized light dull grey purple triangle)
hell-lila Dreieck *(light-purple triangle)*

Figure 6.8: Object marked in scene and example descriptions by participants (images adapted from Mast et al., 2016).

6.3.2 Property Models

The following property models were used (the number of the model in Table 4.1, page 95 is given in brackets):

- full colour model with 37 colour terms (model 6, see also Section 4.4)
- contour based shape model with the concepts SQUARE, TRIANGLE, DISC (model 8, see also Section 4.5)
- size model with concepts LARGE and SMALL, using local prototypes (model 4, see also Section 4.3)
- projective relation model, treating horizontal and vertical relations as equivalent (model 11, see also Section 4.6)
- spatial region model, treating horizontal and vertical regions as equivalent (model 12, see also Section 4.7)

For all models, in addition to the vague model, a corresponding crisp model was created by performing Voronoi tessellation on the acceptability values, as described in Section 4.8.

6.3.3 Parsing

Descriptions showed a large variation regarding length, attribute selection, and – to a smaller degree – category assignment. Figure 6.8 gives an example

of the potential for variability.

A very simple parsing procedure was used to retrieve semantic representations from the provided descriptions. Using a dictionary file, all terms covered by the property models were detected and normalised to a standard form, retaining their original order. In order to discriminate between spatial regions (e.g., *the front circle* vs. *the circle in front of the square*), multi-word units were considered.

All parts of the utterance that could not be identified were removed. This included hedges (e.g., *rötlich (reddish)* being parsed as RED), precision markers (e.g., *ganz rechts (all the way to the right)* being parsed as RIGHT), and modifiers which were considered negligible (e.g., *feuerrot (fire-red)* was parsed as RED). Further, in some cases entire words or properties were unknown and thus ignored (e.g., *der beige Kreis (the beige circle)* being parsed as CIRCLE, ignoring the expression *beige*).

The resulting standardised descriptions were parsed linearly, adding all properties to an object until a spatial relation was reached. Properties following a spatial relation were treated as applying to the reference object, while the target was always assumed to be the first object. Thus the utterance *violettes Dreieck unterhalb von hellrotem Dreieck und rechts neben grauem Kreis (purple triangle below light-red triangle and to the right of grey circle, ID:186)* would be interpreted as {TRIANGLE(x_1), PURPLE(x_1), BELOW(x_1,x_2), LIGHT-RED(x_2), TRIANGLE(x_2) RIGHTOF(x_1,x_3), GREY(x_3), CIRCLE(x_3)}. One object could be ascribed several properties of the same domain, e.g. *kleiner grau lilaner kreis (small **grey purple** circle, ID:124)*, *das kleine Dreieck oben rechts (the small **upper right** triangle, ID:11)*.

The parsed utterances were manually assigned three categories: utterances for which one or more simple properties were not recognised while the rest of the description was still parsed correctly were categorised as *omissions*. This group consisted of 61 utterances (9.8%).

In 43 cases (6.92%), unrecognised relations yielded completely useless representations, e.g. *hellroter Kreis zwischen dem grauen und dem grünen Dreieck (light-red circle **between** the grey and the green triangle)* becoming {LIGHT-RED(x_1), CIRCLE(x_1), GREY(x_1), GREEN(x_1), TRIANGLE(x_1)}

due to *between* not being covered by the property models. Such cases were categorised as *errors*.

The remaining 517 utterances were correctly parsed and fully handled by the model with the exception of hedges, precision markers, and negligible modifiers (see above). These utterances were categorised as *correctly parsed*.

6.3.4 Interpretation and Evaluation

The semantic representations received were input to PRAGR which acted as a listener. As in the robot-robot evaluation in Section 6.2, the listener used the Amy-photos which caused perceptual deviation due to the perspective mismatch, in addition to the human-robot conceptual deviation. Processing of the image and property modelling based on received scenes was conducted exactly as in the robot-robot evaluation. The system performance was then evaluated for (a) all utterances, (b) correctly parsed and omissions only, (c) correctly parsed only. Different rotation adjustments were tested. For this evaluation study, α was ignored, as it is only used in generation.

6.3.5 Results

The results of the evaluation are shown in Table 6.2. Overall, results were robust to perspective mismatch, yielding a maximal difference in success rate of ca. 2 percentage points over the range of tested rotation adjustments. Therefore, Figure 6.9 summarises performance averaged over all rotation adjustments. Figure 6.9a shows the percentage of correct interpretations and Figure 6.9b shows the percentage of cases in which the correct target was among the first two candidates proposed by the system.

Out of all 621 human descriptions, the crisp listener on average correctly identified 64.95%, while vague PRAGR reached a success rate of 75.95%. With an average number of 6.81 objects per image, and given that each image was seen once, the chance of identifying the correct target by random selection lies at 14.68%. Considering only the correct parses, crisp PRAGR achieved 71.63% correct identifications, while vague PRAGR reached 83.05%.

Table 6.2: Results of system's interpretation of human referential utterances

dataset	vagueness	rotation [°]	correct	correct in first two	total	success [%]	correct in first two [%]
all	crisp	20	399	489	621	64.25	78.74
all	crisp	35	403	488	621	64.90	78.58
all	crisp	50	408	488	621	65.70	78.58
all	vague	20	465	553	621	74.88	89.05
all	vague	35	478	560	621	76.97	90.18
all	vague	50	472	556	621	76.01	89.53
omissions	crisp	20	390	469	578	67.47	81.14
omissions	crisp	35	395	470	578	68.34	81.31
omissions	crisp	50	401	469	578	69.38	81.14
omissions	vague	20	451	529	578	78.03	91.52
omissions	vague	35	463	536	578	80.10	92.73
omissions	vague	50	458	532	578	79.24	92.04
clean	crisp	20	366	427	517	70.79	82.59
clean	crisp	35	371	427	517	71.76	82.59
clean	crisp	50	374	427	517	72.34	82.59
clean	vague	20	423	486	517	81.82	94.00
clean	vague	35	434	491	517	83.95	94.97
clean	vague	50	431	488	517	83.37	94.39

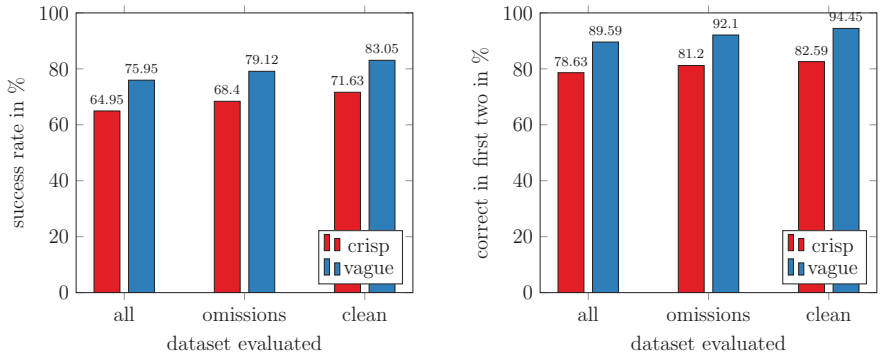


Figure 6.9: Results for interpretation of human descriptions using vague vs. crisp property models, using different test sets. Results are averaged over all rotation adjustments. (a) Percentage of correct reference resolution by system. (b) Percentage of correct target contained in first two guesses of the system.

Throughout all settings, the success rates of vague PRAGR are ca. 10 percentage points above the crisp system. This difference persists when checking whether the correct target is among the best two guesses of the system, where the crisp system achieves 78.63% (all data) to 82.59% (correct parses only), while vague PRAGR achieves 89.59% (all data) to 94.45% (correct parses only).

6.3.6 Discussion

These results confirm that the superiority of vague models shown in the robot-robot interaction also holds for understanding human REs, and is in line with the finding that a vague listener is particularly important for improving referential success.

The results of the vague model compare favourably to those achieved by Gorniak and Roy (2004) whose system achieves an accuracy of 72.5% for a test scenario comparable in difficulty, using only fully parsable descriptions – as compared to PRAGR's 83.05%. As in the work by Gorniak and Roy (2004), the scenario used here has a variety of properties and only few easily describable reference objects. While the scenario by Gorniak and Roy (2004) has more complex spatial configurations, they have highly simplified all other properties. With 13%, the chance performance of their scenario is comparable to the one used here, although the variation of scene complexity is higher in their scenario.

6.4 Evaluating the Understandability of System-Generated Utterances

In order to evaluate PRAGR's success at generating REs that are understandable for humans, I performed a further evaluation study. PRAGR took the role of a speaker and generated REs for objects in simple, automatically generated scenes. In a two-part experiment, human subjects first interpreted the REs generated by the system, and then evaluated the descriptions.

Overall, the implementation used for this experimental setup consists of the following components:

- a scene creation script,
- several property models,
- one instance of the PRAGR reference handler (speaker),

- experimental software for collecting human interpretation and evaluation data, and
- an evaluation script.

In the following, I will describe the creation of the stimuli, the property models used, and the experimental procedure. I will then proceed to present the results and discuss the implications of the experiment.

6.4.1 Stimuli

42 test scenes showing either 5 or 9 rectangles were automatically generated (see Figure 6.10b). Rectangles varied in size and width-height ratio, and were rotated at random angles. Colours were randomly chosen out of the range of fully saturated colours.

6.4.2 Property Models

The property models were adapted in order to suit the conditions of the stimuli. The following property models were used (the number of the model in Table 4.1, page 95 is given in brackets):

- simplified colour model with 8 rainbow colour terms (model 7, see also Section 4.4.3);
- simple shape model with the concepts LONG and SQUARE (model 5, see also Section 4.5.3);
- size model with the concepts LARGE and SMALL, using global prototypes (model 3, see also Section 4.3);
- projective relation model using vertical relations (model 10, see also Section 4.6);

For all models, in addition to the vague model, a corresponding crisp model was created by performing Voronoi tessellation on the acceptability values, as described in Section 4.8.

which case they were counted as incorrect.

In the second part of the experiment, participants evaluated the descriptions according to description quality and uniqueness. For this purpose, participants saw the scene with the target object marked with a black X, next to the evaluation statements, as shown in Figure 6.11. The evaluation statements were phrased as follows:

- Das ist eine gute Beschreibung für dieses Objekt. (That is a good description for this object.)

- Die Beschreibung ist eindeutig. Es besteht keine Gefahr, das beschriebene Objekt mit einem anderen Objekt zu verwechseln. (The description is unambiguous. There is no danger of confusing the described object with another object.)

For each evaluation statement, a sliding scale was provided on which participants could mark their evaluation of how well the evaluation statement holds for the description in question. The extremes of the scale were labelled *trifft gar nicht zu (is not true at all)* and *trifft voll und ganz zu (is absolutely true)*. Once they were satisfied with their evaluation, they pressed a button to proceed to the next scene. Participants evaluated the same pairs of scene and description as they had seen in the identification task. The same scenes were chosen for both tasks in order to allow qualitative evaluation of individual descriptions based on both task success measures and subjective evaluation metrics, while keeping the total number of participants low for reasons of practicality. The evaluation task was started after *all* items of the identification task had been completed, as seeing the scene with the correct target object marked before performing the identification task would strongly influence the results. The influence of identification on the later task of subjective evaluation is less problematic, although it does warrant caution with quantitative evaluation. From a qualitative perspective, one may even argue that the prior experience of having to identify the target based on the description would make participants more aware of the shortcomings of certain descriptions, and thus encourage them to evaluate the descriptions more

Table 6.3: Human identification success for descriptions generated with PRAGR using crisp and vague properties.

	crisp	vague $\alpha = 0.3$	vague $\alpha = 0.35$
correct	451	463	459
total	532	532	532
percent.correct	85	87	86

critically. Overall, this procedure was deemed a reasonable compromise in the face of limited time and participants.

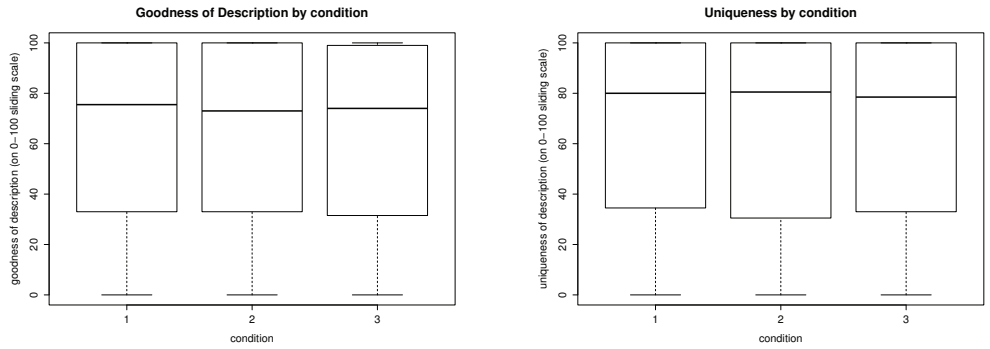
6.4.4 Results

Table 6.3 summarises the results of the identification task. As the table shows, the participants were able to interpret most descriptions correctly, with vague PRAGR performing only marginally better than the crisp version. Percentage of correct interpretations was 85% for crisp PRAGR, and 86% and 87% for the two vague versions. Table 6.4 below shows examples of scenes and the respective descriptions created for each scene by crisp and vague PRAGR.

Figure 6.12 shows boxplots of the subjective human evaluation data. As the plots show, subjective evaluation does not differ between the conditions, and there is strong variation in all conditions. Evaluation of description quality (Figure 6.12a) yields mean values of 64.94 (crisp), 65.34 (vague, $\alpha = 0.3$), and 63.98 (vague, $\alpha = 0.35$), with standard deviations of 34.16 (crisp), 35.30 (vague, $\alpha = 0.3$), and 35.18 (vague, $\alpha = 0.35$). We can conclude that the quality of descriptions varies much more strongly within each condition than between conditions.

6.4.5 Discussion

The fact that human understanding of the generated utterances did not differ between conditions is in line with the results from the robot-robot experiment which show a much stronger benefit of vague categories for the listener, as compared to the speaker. However, it may also be due to the fact that in most scenes the target objects could be easily described using only two



(a) (b)

Figure 6.12: Results for human subjective evaluation of REs generated by PRAGR. (a) Goodness of description. (b) Uniqueness of description.

properties, and in many cases crisp and vague speakers produced exactly the same descriptions (14 descriptions are identical between crisp and vague with $\alpha = 0.3$ and 10 between crisp and $\alpha = 0.35$). Table 6.4 shows four examples in which either the crisp or the vague PRAGR performed particularly badly. In example 1, the vague PRAGR generated a better description than the crisp version. Here, the target (in the bottom right corner) was described as *large yellow object* by the crisp system, probably due to the fact that a minimal difference in size lead to the categorisation of the target as LARGE, and the most relevant distractor as SMALL. However, the distractor object is a slightly better YELLOW than the target, rendering the description confusing. As the target is only minimally larger, the property LARGE does not amend this confusion. The vague generator, on the other hand, selects the small blue square as reference object, as its highly discriminatory features allow a unique description, making it easier to describe the intended target. Similarly, in example 2, the crisp PRAGR ignores the possibility of confusing the target object (object 6) with the distractor object 9. The vague PRAGR, on the other hand, further qualifies the description with those properties which discriminate the target object from the most similar distractor: SMALL, and LONG.

On the other hand, in some cases the vague system produced descrip-

Nr	scene	descriptions
1		

ject (object 5) was described as *square purple object* despite the fact that PURPLE is such a bad fit for this object that most humans would not be willing to follow this conceptualisation at all. This description was selected by vague PRAGR over the better fitting BLUE due to its ability to better distinguish the object from the distractor (object 4) which is also BLUE, but more towards the turquoise spectrum.

While one might argue for solving this issue by increasing α to give more weight to Acceptability, this comes with its own risks, as shown by example 4 where vague PRAGR produces the overly short description *the orange object* despite it not being discriminating. Here, clearly the small blue object would have made for an excellent reference object (and was used as such by crisp PRAGR). However, due to the large distance the Acceptability of this description would have been low. Due to the other blueish distractors, the gain in Discriminatory Power was not sufficient to justify the decrease in Acceptability, yielding a suboptimal description.

It is therefore necessary to explore the relationship between acceptability and Discriminatory Power in more detail, in order to make full use of the possibilities gradedness offers for REG.

6.5 Summary

In this chapter, after a brief discussion of prior evaluation challenges for REG and metrics used for evaluating REG, I presented three empirical studies evaluating the performance of PRAGR in scenarios involving robot-robot and human-robot interaction. With these studies, I tested the ability of PRAGR to produce and understand REs for simple visual scenes under conditions of perceptual deviation, using a number of different property models. In particular, I evaluated the impact of using vague property models, as compared to crisp ones on the generation and resolution of REs.

The studies demonstrated that a probabilistic approach to reference handling with vague properties can help bridge the gap between human and artificial communicators in situated interaction by using flexible concept assignment based on vague property models and situational context with the

goal of maximising the chance of communicative success. The results further showed that PRAGR is capable of understanding human-produced REs with a high degree of accuracy under conditions of perceptual deviation, and can generate REs which are easily understood by human subjects. Further, the evaluation showed that using vague property models improves task success in robot-robot and human-robot communication under conditions of perceptual deviation, in particular on the side of the listener, while a speaker using vague properties showed only limited improvement over crisp categories.

Chapter 7

PRAGR in Grounding Dialogues

As discussed in Section 2.3.3, successful referential communication requires bridging a conceptual gap, mediating via linguistic symbols between the individual semiotic networks of the interactants. This process requires the interactants to flexibly adapt the use of concepts (Clark and Brennan, 1991; Clark and Wilkes-Gibbs, 1986; Garrod and Anderson, 1987). Due to the danger of miscommunication, it cannot be dealt with simply by a speaker unilaterally uttering an RE and a listener resolving that expression. Reference is instead a collaborative process which requires agreement (Clark and Wilkes-Gibbs, 1986), and which forms part of a continuous process of grounding with the goal of ensuring mutual comprehension (Garrod and Anderson, 1987).

In this chapter, I will discuss how PRAGR can be integrated with a dialogue component to allow intelligent referential grounding dialogues. In Section 7.1, I will describe in general terms the ways in which PRAGR can support referential grounding dialogues. In Section 7.2, I will then present the integration of PRAGR with the DAISIE dialogue system framework and architecture (Ross and Bateman, 2009) for a simple referential grounding dialogue scenario.

7.1 Mediating between Perceptual and Dialogic Grounding

In the introduction to this thesis, I presented a short dialogue between Mary and Amanda which I will repeat here for the sake of clarity: Mary is resting on the sofa, relaxing after a long day at work. She decides to read the new book she got for her birthday last week, and asks her personal assistant robot Amanda: *Could you pass me that yellow book on my desk?* With the integrated camera, Amanda scans the desk and identifies that there are two books which seem to match that description. She asks: *Do you mean the one in front of the coffee cup?* Slightly annoyed, Mary replies: *No, not the green one, the yellow one.* Amanda understands and confirms: *Oh, okay. I'll get it.*, moves to the desk, grabs the book, and brings it to Mary.

In this dialogue, Mary and Amanda collaborate on achieving grounding: Mary first provides an RE, *the yellow book on my desk*, and Amanda provides an RE of her own, *the one in front of the coffee cup*, in order to confirm that they are talking about the same object. Mary detects the miscommunication, and provides an overt correction, *not the green one, the yellow one*, which allows Amanda to adapt her construal and identify the originally intended referent. Amanda then confirms that she considers the referential dialogue to have reached a successful ending.

While to my knowledge there is no empirical data on the frequency of this type of grounding dialogue in natural human communication, Conversation Analysts have presented numerous examples of grounding dialogues in experimental settings which allow for free conversation between participants and in natural situations (for an overview see Clark and Bangerter, 2004). These examples include implicit strategies, for example when a signal of non-comprehension from the listener prompts the speaker to elaborate on a prior reference, and explicit negotiation where one partner corrects a noun-phrase proposed by the other. Pickering and Garrod (2004) argue that the norm in communication is for humans to automatically align on various linguistic levels, while explicit negotiation only occurs in cases of misalignment that cannot be easily remedied by more implicit processes. However, they note

that reformulations are in fact very common in everyday communication.

With PRAGR, I have presented a mechanism for handling reference which reflects the fact that calling a book *yellow* or *green* is to some degree a matter of strategic choice – a momentary, strategic decision to construe an object or property in a certain way, in order to improve chances of communicative success.

Firstly, the representation of objects in terms of the acceptability of concepts, rather than a priori conceptualisations allows PRAGR to make situation-specific decisions regarding the conceptualisation of objects when generating REs, and to evaluate the likelihood of candidate referents based on the acceptability of conceptualisations, without the need to commit to any specific conceptualisation of these candidates.

This allows the integration of image understanding (object detection and modelling of qualitative properties) on the one hand, and the planning of dialogue contributions on the other hand. Here, PRAGR has the function of reducing complex perceptual processes to simple quantitative representations (Acceptability, Discriminatory Power, and Appropriateness) which contain the information relevant for decision making in dialogue.

Further, by allowing the evaluation of Acceptability, Discriminatory Power, and Appropriateness of descriptions both in the context of REG and RR, PRAGR provides a strong link between REG and RR. Maximisation of Appropriateness of an RE by the speaker can be used for selecting the best RE, while in RR, low values of Discriminatory Power or Acceptability may be used as information for motivating grounding moves in dialogue.

Using the sub-symbolic meta-information PRAGR provides enables reaction to, and dialogue about mismatched models of the world in a much more subtle way than a system based on crisp properties. For example, if the user commands the robot to *go to the large green box*, a robot sensitive to the information underlying categorisation may be able to identify the correct referent, even if its own preferred conceptualisation of the object would have been *yellow*.

Further, if the sensory information warrants *green* as an acceptable categorisation, this mismatch might be corrected without any further negoti-

ation having the robot respond with *okay, I'll go to the large green box.* While, if the mismatch were fairly large, the robot might react with a clarification question using additional features which, in the given situation, promise more certain categorisation: *do you mean the large box to the left of the red ball?*

7.2 Referential Grounding with DAISIE and PRAGR[1]

As part of this thesis, I implemented a version of the PRAGR reference handler which interfaces with an implementation of the agent based dialogue system architecture and framework DAISIE (Ross and Bateman, 2009) in order to allow referential grounding dialogues. DAISIE features flexible dialogue management with a formal unified dialogue modelling approach combining information state update theories with generalised dialogue models (Shi et al., 2011), a CCG (Steedman and Baldridge, 2011) parser and a KPML (Bateman, 1996) natural language generation module.

While the integration of PRAGR into the DAISIE framework is not the focus of this thesis, I will briefly present at this point the interfacing components, and an example of a decision procedure for handling grounding dialogues with PRAGR.

I will demonstrate based on an example scenario, how the DAISIE + PRAGR system is capable of engaging in grounding dialogues about photographs as they may be provided by a camera installed on the head of a mobile robot, by generating and resolving REs, and using probabilistic evaluations of the REG and RR output for making reasonable dialogue decisions.

[1]The integration of PRAGR with DAISIE was joint work with Daniel Couto Vale, Zoe Falomir, and Mohammed Fazleh Elahi (Mast et al., 2014a).

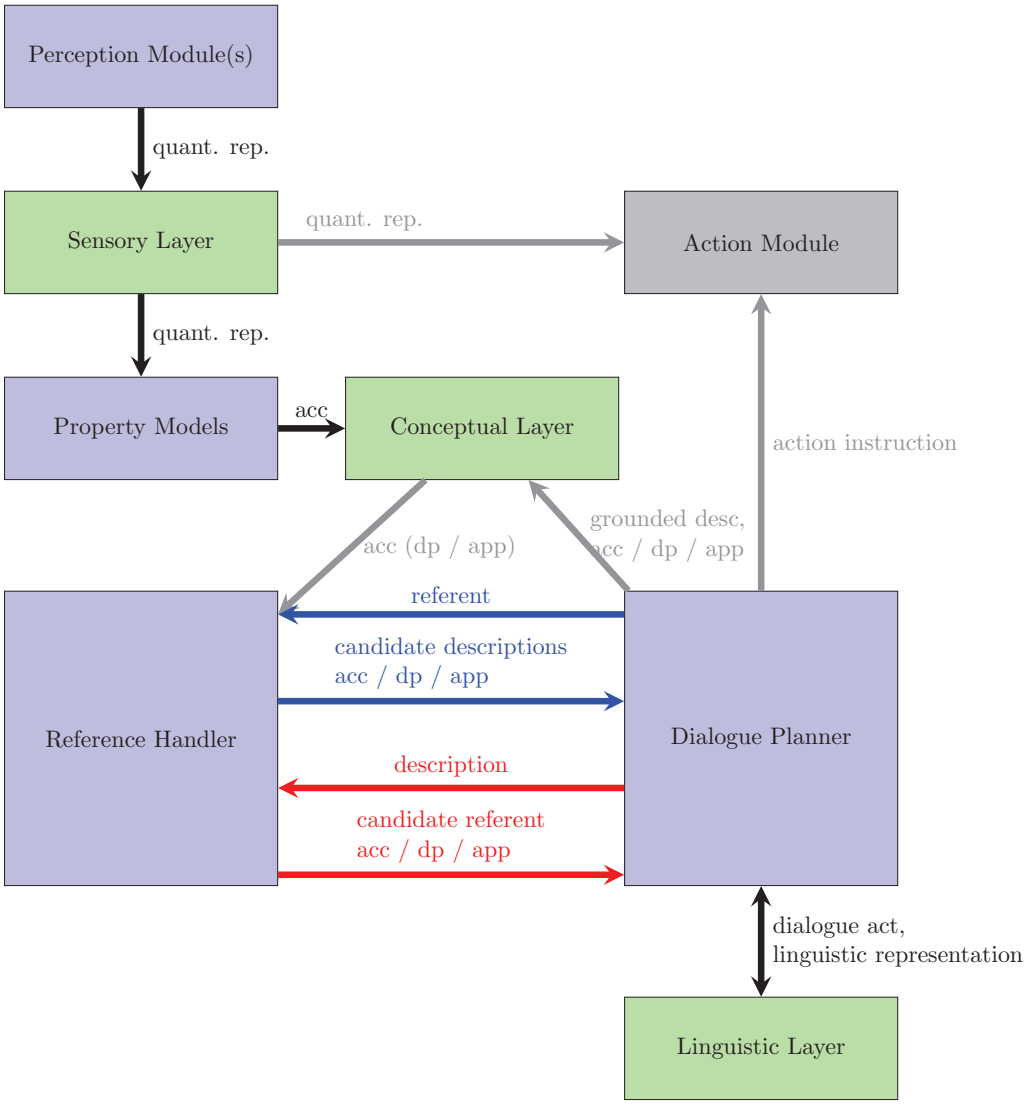


Figure 7.1: Information flow in human-robot referential dialogue with PRAGR.

7.2.1 Layered Representation

Figure 7.1 shows an overview of the flow of information in a human-robot referential dialogue with PRAGR. Knowledge Representation Layers are coloured in green, Processing Modules are coloured in blue. Greyed out Modules and connections have not been implemented.

Note that the Dialogue Planner Module is represented here only in a

massively simplified manner, as this is not the subject of the present thesis.

In its most recent implementation, DAISIE+PRAGR features three layers of representation, realised via separate ontologies: a sensory layer, a conceptual layer, and a linguistic layer. These layers of representation serve as interfaces for different modules to communicate information.

The sensory layer represents information about individual objects on the sensory level, for example the edge points of an object, its height, and representative colour points with associated values of hue, saturation and lightness. The sensory layer is fed by an image segmentation module, and thus relies entirely on primary sensory information. The information of the sensory layer is used by the property models, but it may also be used by an action module, e.g., when the robot has received the instruction to go to a given object, the motion handling module may use the information in the perceptual layer in order to determine the position of the object and of possible obstacles.

The abstraction layer allows representing objects in terms of qualitative properties such as colour (ORANGE or YELLOW) or height (TALL, SHORT) and relations such as projective relations (BEHIND, IN FRONT OF). In order to allow for the integration of probabilistic property modeling, as used by PRAGR, an additional abstraction layer is introduced which allows the creation of mappings between objects, simple or complex qualitative descriptions, and their corresponding acceptability values via description elements. A mapping exists between individual objects on the sensory level, and their counterparts on the conceptual level.

The abstraction layer can be fed by different sources. The primary source is the cognitively motivated property models described in Section 4 which perform a probabilistic mapping from sensory information to qualitative concepts. However, the abstraction layer may also be fed by the dialogue component: Once a (complex or simple) description for a given object has been dialogically grounded, the corresponding mapping in the abstraction level may be added or overwritten by the dialogue component, thus giving dialogically grounded descriptions and properties preference over perceptually grounded ones.

The linguistic layer allows representing the meaning of utterances based

on the theory of Systemic Functional Linguistics (Halliday and Matthiessen, 2004). While the sensory and abstraction layer are concerned only with the state of affairs in the domain, the linguistic layer also includes the interpersonal metafunction which is concerned with the social interaction of the speaker, or what a person is trying to do with their utterance (Couto Vale and Mast, 2012, 2013). For example, the utterance *Bring me the red ball.* can be classified as *Mandative*, as it is a demand of the speaker for the listener to provide some object or resource to them, in this case the referent of *the red ball* (Couto Vale and Mast, 2013).

Thus, the Dialogue Planner processes input information received via the linguistic layer. Based on the dialogue act and a linguistic representation of the received utterance, the Dialogue Planner decides whether a referent needs to be resolved, or whether a description for an object needs to be generated.

If a referent needs to be resolved, a conceptual representation is extracted from the linguistic meaning, and (if necessary) enriched via co-reference resolution against the discourse history during the experiential interpretation before being handed over to the Reference Handler. Thus, in the following dialogue: H: *Bring me the red box.*, S: *Which red box do you mean?*, H: *The one on the floor.*, *one* is resolved as co-referring with *the red box*, and the conceptual representation of the enriched description – RED, BOX, ON(FLOOR) – is passed to the reference handler for resolution within the scene model.

Based on information from the Conceptual Layer, the Reference Handler then produces a list of candidate referents, ordered by the acceptability of the description for the respective referent, enriched by additional information about Discriminatory Power and Appropriateness of the description.

If a description needs to be generated, the Reference Handler is called with the given referent. Based on the information from the Conceptual Layer, the Reference Handler proceeds to generate a list of candidate descriptions for the referent, ordered by Appropriateness.

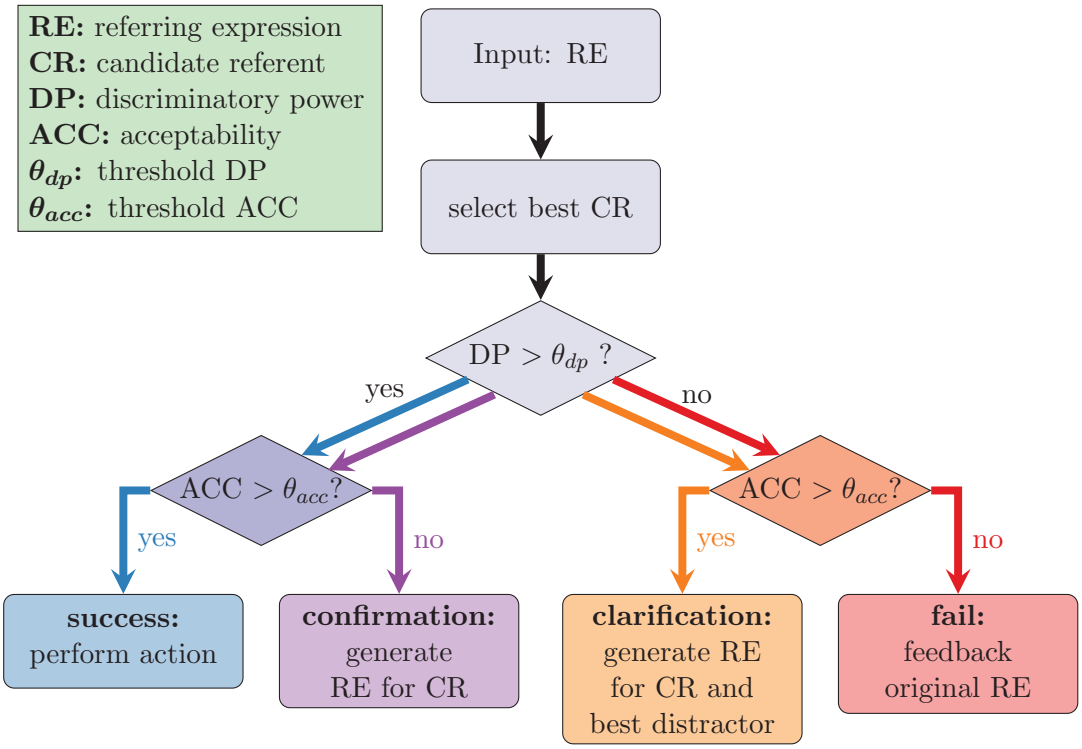


Figure 7.2: Decision procedure for dialogue move.

7.2.2 Simple Grounding Dialogues

A working prototype was implemented for simple grounding dialogues where a human asks a robot to move to a certain object, and the robot engages in grounding behaviour in order to ensure correct understanding of the goal of the motion. Based on an n-best list of potential referents with Acceptability, Discriminatory Power and Appropriateness values of the input description for each, the dialogue manager then makes a decision about the next dialogue move. Figure 7.2 shows the decision making procedure which takes into consideration the values of Acceptability and Appropriateness in order to determine the next dialogue contribution in a referential grounding dialogue. Figure 7.3 shows an example dialogue for each possible system decision.

The dialogue planner receives an n-best list of potential referents sorted by Acceptability. If the Discriminatory Power of the description for the best candidate is higher than a predetermined threshold value, it is highly

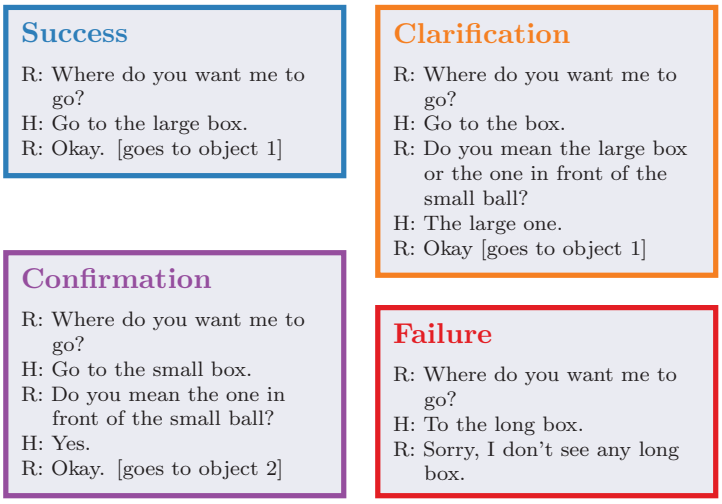


Figure 7.3: Example grounding dialogues with DAISIE+PRAGR.

likely that the correct referent has been identified. If the Acceptability of the description for the candidate is also above a threshold, the planner can safely assume that the correct referent was found and proceed directly to executing the desired task and uttering a positive feedback (blue path in Figure 7.2, blue example in Figure 7.3). If, on the other hand, the Acceptability of the description for the most likely target is below the threshold, it is still likely that the correct object was identified (due to the high Discriminatory Power) but the description itself does not meet the grounding criterion. Thus, the planner calls the Reference Handler again, this time in order to generate the best RE for the assumed referent, and uses this for uttering a confirmation question (purple path in Figure 7.2, purple example in Figure 7.3).

On the other hand, if the Discriminatory Power is lower than the threshold, the given information is not sufficient for confidently selecting a candidate. In this case, if the acceptability for the best candidate is high, this means that the ambiguity of the utterance was high, yielding several good candidate referents. In this case, the Dialogue Planner proceeds by asking a clarification question. The Reference Handler is called to generate an RE for each of the best two candidates, and the system produces an either-or question (orange path in Figure 7.2, orange example in Figure 7.3).

If, on the other hand, the acceptability for the best candidate is below

the threshold, this implies that no good candidate could be found, and the grounding attempt has failed. In this case, the Dialogue Planner generates negative feedback about the description which failed to yield a suitable referent (red path in Figure 7.2, red example in Figure 7.3.

This decision procedure is obviously highly simplified. Some thoughts on how PRAGR could be put to use in a more sophisticated grounding dialogue agent will be discussed in Section 8.3.6.

7.3 Summary

In this chapter I have shown how PRAGR can be integrated with a dialogue component to allow intelligent referential grounding dialogues. I described in general terms the ways in which PRAGR can support referential grounding dialogues and presented the integration of PRAGR with the DAISIE dialogue system framework and architecture (Ross and Bateman, 2009) for a simple referential grounding dialogue scenario.

The simple dialogue examples presented here demonstrate how the probabilistic concepts of Acceptability and Discriminatory Power allow the control of referential grounding dialogues and thus can enable a robot to interact intelligently with a human.

Chapter 8

Conclusion and Outlook

In this final chapter, I will provide a summary of the work presented in this thesis and discuss the main contributions of this work. I will then proceed to outline directions for future work, before ending with some closing remarks.

8.1 Summary

In this thesis, I addressed the challenges of reference in situated human-machine interaction and presented an integrated reference handling mechanism for REG and RR which uses a graded, probabilistic concept of Discriminatory Power in order to enable referential grounding dialogues in the face of perceptual deviation.

In **Chapter 2**, I argued for a perspective on reference which focuses on the nature of language as a tool to collaboratively bridge gaps between individual conceptualisations. I further identified a number of key aspects which an integrated mechanism for reference handling should be able to tackle, most notably graded properties, spatial relations, and salience.

Based on this foundation, in **Chapter 3** I argued for an approach to reference which incorporates vagueness as a fundamental characteristic of categorisation rather than an exceptional case, and argued for a probabilistic generalisation of the concept of Discriminatory Power to vague categories. I presented the core contribution of this thesis, the **P**robabilistic **R**eference

And **GR**ounding mechanism, which follows the Modular Decision Approach (MDA) in allowing for an evaluation of complex REs in a given situational context on the basis of independent property models for different conceptual domains via a combinatory mechanism for evaluating Discriminatory Power and Appropriateness. In an example-based evaluation, I showed that PRAGR mimics a number of empirical findings regarding REG, in particular the preference for more discriminatory attributes, and the ability to produce descriptions using combinations of properties which are not pareto-optimal.

In **Chapter 4**, I presented the approach of Conceptual Spaces (Gärdenfors, 2004b) for modelling complex conceptual domains and introduced all property models that were subsequently used in the thesis, covering the following domains: graded adjectives, colour, shape, projective relations, and projective regions.

In **Chapter 5**, I described the extension of PRAGR to handle some of the challenges for REG in realistic scenarios which were identified in Chapter 2. In particular, I presented the extension of PRAGR for integrating spatial relations into REG, arguing for an approach which considers REG to be supporting visual search, thus requiring the reference object to be identified as a prerequisite for identifying the target object. In this context, I presented an extension of PRAGR to handle salience, demonstrating that this meshes naturally with the overall structure of PRAGR, as salience can be modelled as the prior probability of an object. I further presented a search algorithm for PRAGR with the goal of overcoming some of the complexity issues raised by combining vagueness and spatial relations in REG.

With the three experiments presented in **Chapter 6**, covering both robot-robot and human-robot communication, I empirically evaluated the PRAGR mechanism, testing its ability to produce and understand REs for simple visual scenes under conditions of perceptual deviation, using several of the property models described earlier in Chapter 4. In particular, I evaluated the impact of using vague property models, as compared to crisp ones on the generation and resolution of REs.

Finally, in **Chapter 7**, I showed some preliminary work integrating the PRAGR mechanism into a dialogue system to enable intelligent referential

grounding dialogues. I presented the integration of PRAGR with the DAISIE dialogue system framework and architecture for a simple referential grounding dialogue scenario.

8.2 Contribution of this Thesis

The core contribution of this thesis is the development and evaluation of an integrated mechanism for handling both REG and RR which incorporates vagueness as a fundamental characteristic, and the extensive exploration of the ability of such a mechanism to handle both a wide range of conceptual domains (colour, shape, size, projective relations, projective regions) and a range of relevant phenomena of reference (relations, salience, REG and RR). While during the time taken to complete this thesis, similar approaches have been developed which follow the MDA and thus treat vagueness as fundamental to reference, to my knowledge the work presented here is the first to provide an integrated mechanism which covers such a breadth of domains and phenomena while providing empirical evaluation for the overall approach.

With respect to the kind of properties covered, most approaches discussed in Section 3.1.4 explicitly consider graded adjectives, and some works consider projective terms in terms of vagueness (e.g., Gorniak and Roy, 2004; Kelleher, 2011; Spranger and Pauw, 2012). Properties such as colour and shape are treated as crisp by almost all approaches (with the exception of Meo et al., 2014; Roy, 2002). Of course most approaches can in some way be extended to treat other properties as vague. For example, the approach of van Deemter (2006) can be extended to treat colour as vague (van Deemter, 2016). However, such an extension is by no means trivial, as it may cause further complications. For example, as discussed in Section 3.1.4, in the Independent Decision Approach, modelling several domains with vague properties may lead to preference of suboptimal properties due to a lack of an integrated measure of comparison. Therefore, the work presented here is unique in its wide coverage of different kinds of conceptual domains, and their integration as properties into a single reference mechanism.

Regarding the reference phenomena covered, most existing approaches are restricted to either REG or RR. Most notably however, the treatment of relations in REG in this thesis is unique in that it treats REG with relations as aiding visual search, and integrates findings from reference object selection research, thus providing the first REG system which is capable of highly sophisticated reference object selection with the influence factors of reference object locatability, search space optimisation, and communication cost, as identified by Barclay and Galton (2008). A more thorough empirical evaluation of this capacity would be desirable in order to gain a deeper understanding of reference object selection for REG.

With numerous example-based evaluations throughout this thesis I have demonstrated that PRAGR mimics empirical findings on reference, such as using the most discriminating property, a preference for more salient reference objects, a preference for reference objects which are in a prototypical relation to the target object, and the impact of salience on the length of REs.

With three empirical evaluation studies, covering both robot-robot and human-robot communication, I have demonstrated that a probabilistic approach to reference handling with vague properties can help bridge the gap between human and artificial communicators in situated interaction by using flexible concept assignment based on vague property models and situational context with the goal of maximising the chance of communicative success. I further showed that PRAGR is capable of understanding human-produced REs with a high degree of accuracy under conditions of perceptual deviation, and can generate REs which are easily understood by human subjects. Further, the evaluation showed that using vague property models improves task success in robot-robot and human-robot communication under conditions of perceptual deviation, in particular on the side of the listener, while a speaker using vague properties showed only limited improvement over crisp categories.

I showed that PRAGR can be used as a basis for enabling referential grounding dialogues by handling both REG and RR based on a single integrated reference handling mechanism based on the same underlying concepts. In particular, the ability of PRAGR to easily produce lists of n-best referents

is useful for referential grounding dialogues.

8.3 Directions for Future Work

While this thesis has focused on developing the core PRAGR mechanism and demonstrating its suitability for reference handling in situated human-machine interaction, there remains potential for further expansion and improvement. In the following, I will discuss the most relevant directions for future work which follow from the foundations laid in this thesis, namely addressing some limitations of the core mechanism identified through evaluation, learning model parameters including the inclusion of an additional parameter representing preferences for individual properties, reference to sets, and finally the application of PRAGR in advanced approaches to grounding dialogues.

8.3.1 Overcoming Limitations of the Core Mechanism

The analysis of utterances generated by PRAGR discussed in Section 6.4 showed that PRAGR has a tendency to overestimate the Discriminatory Power of marginally acceptable properties, leading to unnatural descriptions. This problem can to some extent be ameliorated by increasing the model parameter α. However, the potential for compensation is limited, as setting α too high leads to the danger of uninformative descriptions, thus leaving room for improvement of the core mechanism.

The reason for this is that discriminatory power is derived via a linear model which contrasts the acceptability of a description for the target object with the summed acceptability of the description for all objects in the scene. This leads to the effect of preferring properties which are marginally acceptable for the target object, but even more marginal or not acceptable at all for all distractors over properties that are highly acceptable for the target object and moderately acceptable for a distractor. For example, let us assume a scene with two objects and two properties with the Acceptability values: $P(d_1|x_1) = 0.9$, $P(d_1|x_2) = 0.5$, $P(d_2|x_1) = 0.01$, and $P(d_2|x_2) = 0.001$. This

will yield a preference for describing x_1 using d_2 rather than the more intuitively appealing d_1, as the acceptability of d_1 for x_1 is only roughly twice as high as for x_2, while the acceptability of d_2 for x_1 is ten times as high as for x_2. Hence, $P(x_1|d_1) = \frac{0.9}{0.9+0.5} = 0.64$ while $P(x_1|d_2) = \frac{0.01}{0.01+0.001} = 0.91$. This issue is particularly problematic when domains such as colour come into play which have many properties, and therefore there is a high chance of finding a marginally acceptable property for an object which has zero or almost zero acceptability for all other objects.

Engonopoulos and Koller (2014) use a log linear model of Discriminatory Power which reduces this problem: $P(x_1|d_1) = \frac{e^{0.9}}{e^{0.9}+e^{0.5}} = 0.60$, while $P(x_1|d_2) = \frac{e^{0.01}}{e^{0.01}+e^{0.001}} = 0.50$.

Spranger (2011) uses an entirely different approach, treating discriminatory power as the difference between the acceptability of the target and the acceptability of the closest distractor. Thus, for evaluating the acceptability of *the red ball*, only the one distractor which has the highest acceptability for the utterance is considered. This acceptability is then subtracted from the acceptability for the target, yielding a Discriminatory Power value. While this approach lacks the benefit of a direct probabilistic interpretation of Discriminatory Power, it is intuitively appealing – when considering whether to call a given object *the red ball*, a single very red and very ball-like distractor would be as problematic as seven of those objects would be when considering only the question of whether or not a listener will be able to identify the target object.

From the probabilistic point of view taken in this thesis, one might also consider the problem to be that the current version of PRAGR does not adequately reflect uncertainty with respect to Acceptability values. While a difference between $acc = 0.9$ and $acc = 0.5$ can most likely be attributed to a relevant difference which is intersubjectively stable, the difference between $acc = 0.01$ and $acc = 0.001$ may be caused by noise in perception or the model, or may not be reflected in the interlocutor's conceptual representation. Therefore, an approach which treats Acceptability not as fixed value, but as a probability distribution (either learnt from data, or modelled as a Gaussian distribution) may overcome the problems reported here. In this case, a Monte

Carlo approach could be used, i.e., drawing samples from this distribution and calculating Discriminatory Power based on these samples.

However, in order to determine the most suitable way to handle Discriminatory Power, a thorough empirical investigation of the different approaches based on human-likeness and task success data would be desirable. This would be interesting both from a cognitive modelling perspective, in order to determine which approach is best suited as a model for human referential behaviour, and from an engineering perspective, in order to determine the approach which generates the most useful descriptions.

8.3.2 Intrinsic Preferences for Properties

A further limitation of PRAGR which has not been addressed in this thesis is that, independently of Discriminatory Power, certain properties are more or less preferred by human subjects. This holds both regarding preferences for different conceptual domains – colour is generally preferred over size (Pechmann, 1987) – and regarding preferences within a conceptual domain – GREEN is generally preferred over BRITISH RACING GREEN (Meo et al., 2014). Meo et al. (2014) present a Bayesian model for generating and resolving colour references which takes into consideration the preference for different terms (termed availability) as well as Discriminatory Power.

It would be worthwhile investigating how such preferences could be integrated into PRAGR. As has been indicated in Section 3.2, the most promising approach to achieve this would be to explicitly separate $P(D_{said})$ from $P(D_{true})$ – two concepts which are currently conflated in PRAGR. However, while Meo et al. (2014) successfully address the issue of learning both Acceptability and availability from a large dataset of colour descriptions, it remains to be shown whether and how availability values can be successfully determined across conceptual domains.

8.3.3 Learning Model Parameters

Relatedly, the question remains of how appropriate model parameters for PRAGR can be determined. While due to the robustness of PRAGR, relying

on hand-crafted property models and model parameters yielded good results for the evaluation studies presented in this thesis, ideally the models should be optimised using machine learning techniques.

While much work on machine learning in REG has been following the Global Decision Approach (GDA) (Engonopoulos and Koller, 2014; Tellex et al., 2014) where parameters for the entire model including features and decisions specific to REG are learnt holistically from training data, the Modular Decision Approach (MDA) combined with psychologically motivated property models with a limited number of parameters suggested here allows the separate optimisation of parameters for individual conceptual domains. Thus, empirical data of Acceptability judgments by human subjects such as the data collected by Sivik and Taft (1994) could be used to fit the dimension weights, prototypes, and sensitivity parameters for individual property models.

While one might criticise that hand-crafting psychologically motivated property models is hard for complex domains and may lead to suboptimal models, the same ultimately holds for the models learnt using the GDA: as Roy (2002) notes, whether one is hand-crafting models or relying entirely on machine learning to establish which of the perceptual features are relevant for a given property, great care needs to be taken to choose the right perceptual features for the domain in question, as badly chosen features will reduce the quality of the learnt models in both cases.

On the other hand, due to the small set of model parameters and the assumption of mutual independence of concepts, parameter fitting can be accomplished with small datasets, a task which would require much larger amounts of data and computing time if one were to make no assumptions about the structure of graded category membership, and/or holistically learn the REG and property models from one dataset.

Finally, a model parameter which requires special attention is the parameter α. While values between $\alpha = 0.1$ and $\alpha = 0.4$ have yielded good results thus far, the ideal value of the weighting parameter is unclear. In principle, this parameter could be learnt using task success measures based on different settings of α. However, the ideal α may vary depending on context and

the kind of properties used, and learning the weighting between Discriminatory Power and acceptability from task success data may be problematic due to the ability of humans to compensate for problematic behaviour of the machine. There is some tentative evidence that human users fair unexpectedly well with unreasonably long descriptions, making the automatic learning of such weighting parameters difficult (Engonopoulos, personal communication).

8.3.4 Improved Heuristic Search Algorithm

As the evaluation in Section 5.2.5 has shown, the search algorithm developed for this thesis was able to overcome some, but not all complexity issues posed by the problem of REG with vague properties and relations. The existing algorithm is well capable of handling scenes with 20 objects, as long as the number of relations in the final description, and the number of individually appropriate properties which are considered in complex descriptions are limited.

However, in order to achieve efficient REG for cluttered scenes with 100 or more objects, a more efficient algorithm is required. In the following, I will discuss 3 possibilities for increasing the efficiency of the REG search algorithm. The first two are simple adaptations of the existing algorithm, the third takes up the idea of incremental generation seen in the Greedy Heuristic Algorithm (GH) and Incremental Algorithm (IA).

Improving Caching Efficiency

A simple improvement of the existing algorithm in terms of efficiency may be achieved by making more thorough use of the cached lists of best descriptions for given sets of target objects and allowed reference objects: Currently, only exactly matching sets of target object and allowed reference objects are considered. However, let us assume a list of best descriptions for the target object 1 with allowed reference objects 2, 3, 4, 5, and 6 had already been created, with results as shown in Table 8.1. Clearly, objects 5 and 6 were not used at all in any of the best descriptions. Thus, if in a step of resolution

requiring the best descriptions for object 1 with allowed reference objects 2, 3, and 4, this list could also be re-used. As the last (i.e., resolved) description does not contain any reference objects, technically even for target object 1 with only allowed reference object 2, the list could be re-used by simply ignoring all descriptions which contain a non-allowed reference object. Only if the resolved description at the end of the list contains a reference object which is not allowed in the new configuration, or if the list was created allowing *less* reference objects than the new configuration, a new search for this resolution step is required.

Table 8.1: Example of a list of best descriptions for a given target object and set of allowed reference objects {2, 3, 4, 5, 6}, where objects 5 and 6 are not used as reference objects.

Pos	Appr	Description
1	0.91	RED(x_1), BALL(x_1), LEFTOF(x_1,x_2), BLUE(x_2), BOOK(x_2)
2	0.88	RED(x_1), BALL(x_1), BEHIND(x_1,x_3), LARGE(x_3), BALL(x_3)
3	0.87	RED(x_1), BALL(x_1), BEHIND(x_1,x_3), LARGE(x_3), GREEN(x_3), BALL(x_3)
4	0.79	RED(x_1), BALL(x_1), LEFTOF(x_1,x_2), BOOK(x_2)
5	0.78	RED(x_1), BALL(x_1), BEHIND(x_1,x_3), GREEN(x_3), BALL(x_3)
6	0.75	RED(x_1), BALL(x_1), RIGHTOF(x_1,x_4), SMALL(x_4), BALL(x_4)
7	0.71	RED(x_1), BALL(x_1)

Restricting Potential Reference Objects

A further option for improving the efficiency of the existing algorithm may be to restrict the number of objects considered as reference objects. We have seen in Section 5.1.4 that objects which have a low salience are unlikely to be chosen as reference objects. This holds both empirically and relating to the way PRAGR evaluates Discriminatory Power. In particular, objects which are less salient than the target object are not very likely to be chosen as

reference objects. Therefore, it may be an option to limit the choice of potential reference objects in the search algorithm to those which are as salient as the target object or more salient. Interpreting this slightly more leniently in order to allow for cases where an object with low salience happens to allow a very easy description, it might be more adequate to allow all objects which are at least half as salient as the target object as potential reference objects (or some similar limit). Alternatively, one might additionally allow those objects which can be described with high appropriateness without the use of relations, as they too are naturally good candidates for reference objects. By limiting the number of potential reference objects, it is effectively possible to limit the search space and retain the processing speed of small scenes when generating REs for larger scenes.

Using a Greedy Approach

Finally, it may be possible to incorporate some incremental generation into the search algorithm while still avoiding the worst pitfalls of that strategy with respect to relations, in particular the issue of recursive dependence.

In this approach, for each resolution step, first the best resolved description (RD) for the sub-target would be generated using a greedy approach by adding in each subsequent step the property which increases Appropriateness the most, until no further property increases Appropriateness any further. Now this part of the description is treated as fixed and combined with possible combinations of the n best relations, and the resulting descriptions are added to the queue as in the original algorithm described in Section 5.2. Again, the best unresolved description (UD) is retrieved from the queue and resolved. By treating the non-relational part of the description greedily, the number of possible descriptions which need to be considered can be massively reduced, while using breadth-first, best first search with respect to the relational properties takes care of the issue of recursive dependence, weeding out relational properties whose reference object cannot be appropriately described and keeping relevant alternatives in the queue.

Once the best description has been found this way, it may be necessary to

perform local optimisation by subsequently testing whether individual non-relational properties can be removed without impacting appropriateness, as they may have been made superfluous by the added relations.

Using a combination of the improvement strategies discussed here, it should be possible to achieve the necessary efficiency to be able to handle scenes with 100 objects and more. However, future work would have to evaluate in detail whether this is indeed the case, and whether the resulting descriptions are good enough approximations of the optimal description to make this worthwhile.

8.3.5 Higher-level Strategies

An aspect of reference which has only been briefly touched by this thesis is that of hierarchical strategies for referring. In an experiment where human subjects were to explain their location within a building to either another human or a supposed computational system (Mast and Bergmann, 2013), participants frequently used a strategy of hierarchically narrowing down the region, starting with the building as a whole, and then narrowing down their position down to the floor, the general area, and the specific position within that area. Similarly, Clarke et al. (2013) show some examples of reference chains starting with a large and highly salient object, and then progressing via a less salient object before identifying the target.

Relatedly, the human data gathered for the evaluation of PRAGR's understanding of human utterances shows a high frequency of spatial regions such as *das Dreieck ganz rechts außen (the triangle right on the outer right)*.

Given the architecture of PRAGR, the first kind of example (hierarchical reference chains) may fall out naturally from applying PRAGR with integrated salience measure to highly cluttered scenes, while the second kind (spatial regions) is currently treated straightforwardly as a unary property with a graded acceptability area. While it seems that PRAGR can handle both kinds of higher-level strategies, this would be worth investigating in more detail, as higher-level strategies can be seen as a key to successful referential communication.

A related issue which has not been addressed in this thesis is reference to sets of objects. Referring to sets of objects is both relevant as a goal in and of itself, as in some cases humans may want to refer to sets of objects, and as a higher-level strategy for achieving successful reference to individual objects by using a set of objects as an reference object.

In principle, it should be straightforward to extend PRAGR to be able to successfully refer to sets of objects which share certain properties that discriminate them from other objects. For such a scenario, one could simply consider both the Acceptability and the Discriminatory Power of the description to groups of n objects, $P(x_1, \ldots, x_n|D)$ which requires determining the Acceptability of a description for multiple objects based on the individual Acceptability for each object.

Further, an extension of PRAGR to reference to sets would allow including second order properties such as *the parallel lines* which hold true of a set, but cannot hold true of an individual object (Stone, 2000).

8.3.6 Advanced Referential Grounding Dialogues

Finally, much research remains to be done with respect to the application of PRAGR in dialogue. While I presented preliminary work on using PRAGR for referential grounding dialogues in Chapter 7, the dialogue strategies presented there are obviously highly simplified. In particular, the dialogue flow described is exclusively concerned with the possibility of conceptual mismatch and does not take into consideration communicative problems on other levels.

Clark (1996) distinguishes four levels on which grounding needs to be achieved for successful communication which Paek and Horvitz (1999) term (1) the channel level, (2) the signal level, (3) the intention level, and (4) the conversation level.

On the channel level, a speaker performs a certain behaviour to which the listener needs to attend to in order for communication to occur. In order for channel level grounding to be successful, the listener needs to be be aware that the speaker's behaviour is directed at them and pay attention

to it (Paek and Horvitz, 1999). On the signal level, the speaker presents a communicative (verbal or non-verbal) signal to the listener. In order for this level of grounding to succeed, the listener needs to correctly identify the signal the speaker produced for them (Paek and Horvitz, 1999). On the intention level, the listener needs to correctly identify the intention the speaker wants to convey with their signal (Paek and Horvitz, 1999). This goes beyond the pure semantic content of the utterance and aligns with Grice's *speaker's meaning* (Grice, 1975). Finally, on the conversation level, the speaker proposes a joint project to the listener who is expected to collaborate in this project to some degree. The listener may take up the joint project or reject it (Paek and Horvitz, 1999).

The different levels of grounding form a ladder where success on a higher level depends on successful grounding at the lower level. This further implies that evidence for success on a higher level can be treated as evidence for success on all lower levels (Clark, 1996).

The dialogue decision procedure presented in this thesis is focused on the intention level, i.e., the question of identifying correctly which object the user wants the robot to move to. Paek and Horvitz (1999) present a Bayesian model for decision making in dialogue which integrates probabilistic information on the likelihood of success at all levels of grounding in order to make decisions about dialogue moves. The probabilistic nature of PRAGR would allow integration into such comprehensive grounding dialogue architectures, allowing the treatment of referential grounding problems as one of several aspects of the entire grounding process.

This would make it possible, for example, to discriminate in a situation with a low probability of correct RR between (a) a case where the likelihood of correct speech recognition is low, while the likelihood of correct visual perception of a scene is high and (b) a case where the likelihood of correct visual perception is low, while the likelihood of correct speech recognition is high. In case (a), the best dialogue move may be to initiate a grounding behaviour on the level of speech recognition, e.g., *I'm not sure I understood you correctly. Did you say 'the large box?'* while in case (b), a response on the conversation level may be more appropriate, e.g., *I can't see the large box,*

where is it? where the robot takes up the user's project and elicits further information in order to be able to contribute the desired action.

8.4 Concluding Remarks

The field of REG has a vibrant community which has been influenced strongly by the early work of Dale and Reiter. Thus, when setting out on the journey of this thesis, this was the natural starting point. Soon I began reacting to limitations I saw in classic REG, and working towards broadening the perspective on reference. My core concern was overcoming the notion of crisp categories which seemed to limit what was possible with REG, in particular with respect to handling spatial relations. During the years it took to complete this thesis, this led to discovery of a range of mostly new approaches which used similar ideas to mine, coming from a different direction – often from a robotics background, driven by concerns of perceptual grounding, sometimes also communicative grounding (Spranger, 2011). Likewise, I studied psycholinguistic research on reference production of humans, and became increasingly interested in the cognitive perspective on reference. Interaction with these works has shaped my perception of reference, and led to a stronger focus on issues of symbol grounding and interaction, which have in turn informed my perspective on handling relations, by leading me to let go of logical definitions of Discriminatory Power, towards aiding visual search.

To some degree, this personal journey also reflects the development of the field, which has become more diverse, and which has seen an increasing amount of interaction between researchers dealing with reference from different angles, for example in the RefNet network on reference[1] which has initiated workshops and publications working towards a computational psycholinguistics of reference (van Deemter et al., 2012b), or in joint workshops of linguists and roboticists on situated language in human-robot interaction such as the Workshop on Spatial Reasoning and Interaction for Real-World

[1] http://homepages.abdn.ac.uk/k.vdeemter/pages/RefNet/index.html

Robotics[2].

Despite leaving many things unsaid and undone, I hope that this thesis has contributed to the growing together of these different fields, and has provided a step towards a broader view of reference that takes into consideration both the physicality of the world with the ensuing difficulties of perception in human and artificial agents, and the collaborative nature of communication and its role in overcoming mismatches in the individuals' perception and conceptualisation of the world.

[2]http://iros2015spatial-workshop.lsr.ei.tum.de/

Bibliography

Abbott, B. (2010). *Reference*. Oxford Surveys in Semantics & Pragmatics. Oxford University Press, Oxford.

Appelt, D. E. (1985). Planning English referring expressions. *Artificial Intelligence*, 26(1):1–33.

Appelt, D. E. and Kronfeld, A. (1987). A computational model of referring. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pages 640–647. AAAI Press / International Joint Conferences on Artificial Intelligence.

Arkin, M., Chew, L. P., Huttenlocher, D. P., Kedem, K., and Mitchell, J. S. B. (1991). An efficiently computable metric for comparing polygonal shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 13:206–209.

Arts, A., Maes, A., Noordman, L., and Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics*, 43(1):361–374.

Barclay, M. and Galton, A. (2008). An influence model for reference object selection in spatially locative phrases. In Freksa, C., Newcombe, N., Gärdenfors, P., and Wölfl, S., editors, *Spatial Cognition VI. Learning, Reasoning, and Talking about Space*, volume 5248 of *LNCS*, pages 216–232. Springer, Berlin/ Heidelberg.

Barclay, M. and Galton, A. (2013). Selection of reference objects for locative expressions: The importance of knowledge and perception. In Tenbrink, T., Wiener, J. M., and Claramunt, C., editors, *Representing Space*

in Cognition: Interrelations of Behaviour, Language, and Formal Models, volume 8 of *Explorations in Language and Space*, pages 59–86. Oxford University Press, Oxford.

Barclay, M. J. (2010). *Reference Object Choice in Spatial Language: Machine and Human Models*. PhD thesis, University of Exeter.

Bateman, J. A. (1996). Kpml development environment: Multilingual linguistic resource development and sentence generation. Technical report, German National Center for Information Technology (GMD), Institute for integrated publication and information systems (IPSI), Darmstadt, Germany. Release 1.0.

Bateman, J. A. (1999). Using aggregation for selecting content when generating referring expressions. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 127–134. Association for Computational Linguistics.

Belz, A. and Gatt, A. (2008). Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *Proceedings of ACL-08: HLT, Short Papers*, pages 197–200. Association for Computational Linguistics.

Belz, A., Kow, E., Viethen, J., and Gatt, A. (2008). The GREC challenge: overview and evaluation results. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 183–193. Association for Computational Linguistics.

Berlin, B. and Kay, P. (1969). *Basic color terms: Their universality and evolution*. University of California Press, Berkeley, CA.

Beun, R.-J. and Cremers, A. (2001). Multimodal reference to objects: An empirical approach. In Bunt, H. and Beun, R.-J., editors, *Cooperative Multimodal Communication*, volume 2155 of *LNCS*, pages 64–86. Springer, Berlin/ Heidelberg.

Bleys, J., Loetzsch, M., Spranger, M., and Steels, L. (2009). The grounded colour naming game. In *Proceedings of Spoken Dialogue and Human-Robot Interaction workshop at the RoMan 2009 conference*.

Blum, H. (1973). Biological shape and visual science. *Journal of Theoretical Biology*, 38(2):205–287.

Bohnet, B. (2008). The fingerprint of human referring expressions and their surface realization with graph transducers. In *Proceedings of the Fifth International Natural Language Generation Conference (INLG)*, pages 207–210. Association for Computational Linguistics.

Bohnet, B. (2009). Generation of referring expression with an individual imprint. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG)*, pages 185–186. Association for Computational Linguistics.

Bohnet, B. and Dale, R. (2005). Viewing referring expression generation as search. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1004–1009, Edinburgh, Scotland. Morgan Kaufmann.

Brainard, D. H. (2003). Color appearance and color difference specification. In Shevell, S. K., editor, *The Science of Color*, pages 191–216. Elsevier Science Ltd, Amsterdam, 2nd edition.

Briscoe, A. D. and Chittka, L. (2001). The evolution of color vision in insects. *Annual review of entomology*, 46(1):471–510.

Bryant, D. J. and Wright, W. G. (1999). How body asymmetries determine accessibility in spatial frameworks. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, 52(2):487–508.

Burnett, G., Smith, D., and May, A. (2001). Supporting the navigation task: Characteristics of 'good' landmarks. *Contemporary Ergonomics*, 1:441–446.

Byron, D., Koller, A., Striegnitz, K., Cassell, J., Dale, R., Moore, J., and Oberlander, J. (2009). Report on the first nlg challenge on generating instructions in virtual environments (give). In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG)*, pages 165–173. Association for Computational Linguistics.

Canny, J. F. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 8:679–697.

Carlson, L. and Covey, E. (2005). How far is near? Inferring distance from spatial descriptions. *Language and Cognitive Processes*, 20(5):617–632.

Carlson, L. and Hill, P. L. (2009). Formulating spatial descriptions across various dialogue contexts. In Coventry, K., Tenbrink, T., and Bateman, J. A., editors, *Spatial Language and Dialogue*, pages 89–103. Oxford University Press, Oxford.

Carlson, L., Regier, T., and Covey, E. (2003). Defining spatial relations: Reconciling axis and vector representations. In van der Zee, E. and Slack, J., editors, *Representing Direction in Language and Space*, pages 111–131. Oxford University Press, Oxford.

Carlson-Radvansky, L. A. and Irwin, D. E. (1993). Frames of reference in vision and language: Where is above? *Cognition*, 46(3):223 – 244.

Carlson-Radvansky, L. A. and Logan, G. D. (1997). The influence of reference frame selection on spatial template construction. *Journal of Memory and Language*, 37(3):411–437.

Clark, H. H. (1996). *Using Language*. Cambridge University Press, Cambridge. Cambridge University Press.

Clark, H. H. and Bangerter, A. (2004). Changing ideas about reference. In Sperber, D. and Noveck, I. A., editors, *Experimental Pragmatics*. Palgrave Macmillan, Hampshire, NY.

Clark, H. H. and Brennan, S. E. (1991). Grounding in communication. *Perspectives on Socially Shared Cognition*, 13:127–149.

Clark, H. H., Schreuder, R., and Buttrick, S. (1983). Common ground at the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior*, 22(2):245–258.

Clark, H. H. and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22(1):1 – 39.

Clarke, A. D. F., Elsner, M., and Rohde, H. (2013). Where's Wally: The influence of visual salience on referring expression generation. *Frontiers in Psychology*, 4(329).

Cohn, A. G. and Gotts, N. M. (1996). The 'egg-yolk' representation of regions with indeterminate boundaries. In Burrough, P. and Frank, A. M., editors, *Proceedings, GISDATA Specialist Meeting on Geographical Objects with Undetermined Boundaries*, pages 171–187, London. Taylor & Francis.

Coleman, L. and Kay, P. (1981). Prototype semantics: The english word lie. *Language*, pages 26–44.

Couto Vale, D. and Mast, V. (2012). Key interpersonal communication skills for wheelchairs. In *Proceedings of IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 421–426.

Couto Vale, D. and Mast, V. (2013). Tacit social contracts for wheelchairs. In *Proceedings of the SIGDIAL 2013 Conference*, pages 294–303, Metz, France. Association for Computational Linguistics.

Crawford, L. E., Regier, T., and Huttenlocher, J. (2000). Linguistic and non-linguistic spatial categorization. *Cognition*, 75(3):209–235.

Dale, R. (1989). Cooking up referring expressions. In *Proceedings of the 27th annual meeting on Association for Computational Linguistics*, pages 68–75, Vancouver. Association for Computational Linguistics.

Dale, R. (1992). *Generating Referring Expressions: Constructing Descriptions in a Domain of Objects and Processes*. MIT Press, Cambridge, MA.

Dale, R. (1995). An introduction to natural language generation. Workshop slides ESSLI Barcelona. retrieved from `http://clt.mq.edu.au/~rdale/teaching/esslli/index.html` 2016-03-22.

Dale, R. and Haddock, N. (1991). Generating referring expressions involving relations. In *Proceedings of the Fifth Meeting of the European Chapter of the Association for Computational Linguistics*, Berlin, Germany.

Dale, R. and Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 18:233–263.

Dale, R. and Viethen, J. (2009). Referring expression generation through attribute-based heuristics. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 58–65. Association for Computational Linguistics.

Di Fabbrizio, G., Stent, A. J., and Bangalore, S. (2008). Referring expression generation using speaker-based attribute selection and trainable realization (attr). In *Proceedings of the Fifth International Natural Language Generation Conference (INLG)*, pages 211–214. Association for Computational Linguistics.

Donnellan, K. S. (1966). Reference and definite descriptions. *The Philosophical Review*, 75(3):281–304.

Dorffner, G. (1992). A step toward sub-symbolic language models without linguistic representations. In Reilly, R. and Sharkey, N., editors, *Connectionist Approaches to Natural Language Processing*, volume 1. Lawrence Erlbaum, New Haven/Hillsdale/Hove.

Douven, I., Decock, L., Dietz, R., and Égré, P. (2013). Vagueness: A conceptual spaces approach. *Journal of Philosophical Logic*, 42(1):137–160.

Engelhardt, P. E., Bailey, K. G., and Ferreira, F. (2006). Do speakers and listeners observe the Gricean maxim of quantity? *Journal of Memory and Language*, 54(4):554 – 573.

Engonopoulos, N. and Koller, A. (2014). Generating effective referring expressions using charts. In *Proceedings of the eigth International Natural Language Generation Conference (INLG)*, pages 6–15, Philadelphia. Association for Computational Linguistics.

Eyre, H. and Lawry, J. (2014). Language games with vague categories and negations. *Adaptive Behavior*, 22(5):289–303.

Falomir, Z., Mast, V., Couto Vale, D., Museros, L., and Gonzalez-Abril, L. (2014). Towards a fuzzy colour descriptor sensitive to the context. In *XVI Jornadas de ARCA – Sistemas Cualitativos y sus Aplicaciones en Diagnosis, Robótica e Inteligencia Ambiental*.

Falomir, Z., Museros, L., and Gonzalez-Abril, L. (2015). A model for colour naming and comparing based on conceptual neighbourhood. An application for comparing art compositions. *Knowledge-Based Systems*, 81:1–21.

Falomir, Z., Museros, L., Gonzalez-Abril, L., and Sanz, I. (2013). A model for qualitative colour comparison using interval distances. *Displays*, 34:250–257.

Frank, M. C. and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Frank, M. C. and Goodman, N. D. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive psychology*, 75:80–96.

Franklin, N. and Tversky, B. (1990). Searching imagined environments. *Journal of Experimental Psychology: General*, 119(1):63–76.

Frege, G. (1892). Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, NF 100(1):25–50.

Frixione, M. and Lieto, A. (2012). Prototypes vs exemplars in concept representation. In *Proceedings of International Conference on Knowledge Engineering and Ontology Development (KEOD)*, pages 226–232.

Funakoshi, K., Nakano, M., Tokunaga, T., and Iida, R. (2012). A unified probabilistic approach to referring expressions. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 237–246. Association for Computational Linguistics.

Gapp, K. (1995a). Object localization: Selection of optimal reference objects. In Frank, A. and Kuhn, W., editors, *Spatial Information Theory A Theoretical Basis for GIS*, volume 988 of *LNCS*, pages 519–536. Springer Berlin/ Heidelberg.

Gapp, K.-P. (1995b). An empirically validated model for computing spatial relations. In *Proceedings of KI-95*. Springer.

Gapp, K.-P. (1996). Selection of best reference objects in object localizations. In *AAAI Technical Report SS-96-03*, pages 23–34.

Gärdenfors, P. (2004a). Conceptual spaces as a framework for knowledge representation. *Mind and Matter*, 2(2):9–27.

Gärdenfors, P. (2004b). *Conceptual Spaces: The Geometry of Thought*. MIT Press, Cambridge, MA.

Gärdenfors, P. (2011). Semantics based on conceptual spaces. In *Proceedings of 4th Indian Conference on Logic and its Applications*, volume 6521 of *LNAI*, pages 1–11, Berlin/ Heidelberg. Springer.

Garrod, S. and Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27:181–218.

Garrod, S. and Doherty, G. (1994). Conversation, co-ordination and convention: An empirical investigation of how groups establish linguistic conventions. *Cognition*, 53(3):181–215.

Gatt, A. (2007). *Generating coherent references to multiple entities.* PhD thesis, University of Aberdeen, UK.

Gatt, A. and Belz, A. (2008). Attribute selection for referring expression generation: New algorithms and evaluation methods. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 50–58. Association for Computational Linguistics.

Gatt, A., Belz, A., and Kow, E. (2008). The TUNA challenge 2008: Overview and evaluation results. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 198–206. Association for Computational Linguistics.

Gatt, A., Belz, A., and Kow, E. (2009). The tuna-reg challenge 2009: Overview and evaluation results. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG)*, pages 174–182. Association for Computational Linguistics.

Gatt, A., van Gompel, R. P., Krahmer, E., and van Deemter, K. (2012). Does domain size impact speech onset time during reference production? In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*, pages 1584–1589.

Gatt, A., van Gompel, R. P. G., van Deemter, K., and Krahmer, E. (2013). Are we bayesian referring expression generators? In van Deemter, K., Gatt, A., van Gompel, R. P. G., and Krahmer, E., editors, *Proceedings of Production of Referring Expressions: Bridging the Gap betweeen Cognitive and Computational Approaches to Reference (PRE-CogSci 2013).*

Goldstone, R. L., Kersten, A., and Carvalho, P. F. (2012). Concepts and categorization. In Healy, A. F. and Proctor, R. W., editors, *Comprehensive Handbook of Psychology, Volume 4: Experimental psychology*, pages 607–630. Wiley, New Jersey.

Golland, D., Liang, P., and Klein, D. (2010). A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 Conference on*

Empirical Methods in Natural Language Processing, pages 410–419. Association for Computational Linguistics.

Gorniak, P. and Roy, D. (2004). Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, 21:429–470.

Gorniak, P. and Roy, D. (2005). Probabilistic grounding of situated speech using plan recognition and reference resolution. In *Proceedings of the 7th International Conference on Multimodal Interfaces*, pages 138–143. ACM.

Gottfried, B. (2008). Qualitative similarity measures—the case of two-dimensional outlines. *Computer Vision and Image Understanding*, 110(1):117–133.

Grice, H. P. (1975). Logic and conversation. In Cole, P. and Morgan, J. L., editors, *Speech Acts*, volume 3 of *Syntax and Semantics*, pages 43–58. Academic Press, New York.

Halliday, M. A. K. and Matthiessen, C. M. I. M. (2004). *An Introduction to Functional Grammar*. Edward Arnold, London, 3rd edition.

Hayward, W. G. and Tarr, M. J. (1995). Spatial language and spatial representation. *Cognition*, 55:39–84.

Haywood, S. L., Pickering, M. J., and Branigan, H. P. (2005). Do speakers avoid ambiguities during dialogue? *Psychological Science*, 16(5):362–366.

Hermann, T. (1990). Vor, hinter, rechts und links: das 6H-Modell. Psychologische Studien zum sprachlichen Lokalisieren/In front of, behind, left and right: the 6H-model. Psychological studies in verbal localisation. *Zeitschrift für Literaturwissenschaft und Linguistik*, 20(78):117–140.

Hermann, T. and Deutsch, W. (1976). *Psychologie der Objektbenennung*. Huber, Bern.

Hermann, T. and Laucht, M. (1976). On multiple verbal codability of objects. *Psychological Research*, 38:355–368.

Hirtle, S. C. and Heidorn, P. B. (1993). The structure of cognitive maps: Representations and processes. *Advances in Psychology*, 96:170–192.

Hirtle, S. C. and Jonides, J. (1985). Evidence of hierarchies in cognitive maps. *Memory and Cognition*, 13(3):208–217.

Hois, J., Tenbrink, T., Ross, R., and Bateman, J. (2009). GUM-Space. The Generalized Upper Model spatial extension: a linguistically-motivated ontology for the semantics of spatial language. Technical report, University of Bremen, SFB/TR8 Spatial Cognition – OntoSpace.

Horacek, H. (1997). An algorithm for generating referential descriptions with flexible interfaces. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 206–213. Association for Computational Linguistics.

Horacek, H. (2004). On referring to sets of objects naturally. In Belz, A., Evans, R., and Piwek, P., editors, *Natural Language Generation*, volume 3123 of *LNCS*, pages 70–79. Springer, Berlin/ Heidelberg.

Horacek, H. (2005). Generating referential descriptions under conditions of uncertainty. In *Proceedings of the 10th European Workshop on Natural Language Generation (ENLG)*, pages 58–67, Aberdeen. Association for Computational Linguistics.

Horton, W. S. and Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, 59(1):91–117.

Huttenlocher, J., Hedges, L. V., and Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review*, 98(3):352.

Itti, L. (2007). Visual salience. *Scholarpedia*, 2(9):3327. Revision nr. 72776.

Johannsen, K. and De Ruiter, J. P. (2013). Reference frame selection in dialog: Priming or preference? *Frontiers in Human Neuroscience*, 7(667).

Jordan, P. and Walker, M. (2000). Learning attribute selections for non-pronominal expressions. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 181–190. Association for Computational Linguistics.

Jordan, P. W. (2000). *Intentional influences on object redescriptions in dialogue: Evidence from an empirical study.* PhD thesis, University of Pittsburgh.

Jordan, P. W. and Walker, M. A. (2005). Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research*, 24:157–194.

Jurafsky, D. and Martin, J. H. (2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Prentice Hall, Upper Saddle River, NJ [u.a.], 2nd edition.

Kamp, H. (1975). Two theories of adjectives. In Keenan, E., editor, *Formal Semantics of Natural Language*, pages 123–155. Cambridge University Press, Cambridge.

Kelleher, J. and Costello, F. (2005). Cognitive representations of projective prepositions. In Kordoni, V. and Villavicencio, A., editors, *Proceedings of the 2nd ACL-Sigsem Workshop on The Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, Colchester, U.K. ACL-Sigsem.

Kelleher, J., Costello, F., and van Genabith, J. (2005). Dynamically structuring, updating and interrelating representations of visual and linguistic discourse context. *Artificial Intelligence*, 167(1):62–102.

Kelleher, J. and Kruijff, G.-J. (2005). A context-dependent algorithm for generating locative expressions in physically situated environments. In Mellish, C., Reiter, E., Jokinen, K., and Wilcock, G., editors, *Proceedings of the 10th European Workshop on Natural Language Generation*, Aberdeen. ACL-SIGGEN.

Kelleher, J. and van Genabith, J. (2004). Visual salience and reference resolution in simulated 3-d environments. *Artificial Intelligence Review*, 21(3-4):253–267.

Kelleher, J. D. (2006). Attention driven reference resolution in multimodal contexts. *Artificial Intelligence Review*, 25(1-2):21–35.

Kelleher, J. D. (2011). Visual salience and the other one. In Chiarcos, C., Claus, B., and Grabski, M., editors, *Salience: Multidisciplinary Perspectives on Its Function in Discourse*, pages 205–228. Walter de Gruyter.

Kelleher, J. D. and Costello, F. J. (2009). Applying computational models of spatial prepositions to visually situated dialog. *Computational Linguistics*, 35(2):271–306.

Kelleher, J. D. and Kruijff, G.-J. M. (2006). Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 1041–1048, Sydney, Australia. Association for Computational Linguistics.

Kelleher, J. D. and van Genabith, J. (2006). A computational model of the referential semantics of projective prepositions. In Saint-Dizier, P., editor, *Syntax and Semantics of Prepositions*. Kluwer.

Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, 30(1):1–45.

Klein, E. (1980). A semantics for positive and comparative adjectives. *Linguistics and Philosophy*, 4(1):1–45.

Klippel, A. and Montello, D. R. (2007). Linguistic and nonlinguistic turn direction concepts. In *Spatial Information Theory*, pages 354–372. Springer, Berlin/ Heidelberg.

Koller, A., Striegnitz, K., Gargett, A., Byron, D., Cassell, J., Dale, R., Moore, J., and Oberlander, J. (2010). Report on the second NLG chal-

lenge on generating instructions in virtual environments (GIVE-2). In *Proceedings of the 6th International Natural Language Generation Conference (INLG)*, pages 243–250. Association for Computational Linguistics.

Koolen, R., Krahmer, E., and Swerts, M. (2015). How distractor objects trigger referential overspecification: Testing the effects of visual clutter and distractor distance. *Cognitive Science*.

Krahmer, E. and Theune, M. (2002). Efficient context-sensitive generation of referring expressions. In van Deemter, K. and Kibble, R., editors, *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*, pages 223–264. CSLI Publications.

Krahmer, E. and van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Krahmer, E., van Erk, S., and Verleg, A. (2003). Graph-based generation of referring expressions. *Computational Linguistics*, 29(1):53–72.

Krauss, R. M. and Weinheimer, S. (1966). Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, 4(3):343–6.

Kronfeld, A. (1990). *Reference and Computation: An Essay in Applied Philosophy*. Cambridge University Press, Cambridge.

Kruijff, G.-J. M., Kelleher, J. D., and Hawes, N. (2006). Information fusion for visual reference resolution in dynamic situated dialogue. In *Perception and Interactive Technologies*, pages 117–128. Springer, Berlin/ Heidelberg.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. The University of Chicago Press, Chicago/ London.

Lassiter, D. (2011). Vagueness as probabilistic linguistic knowledge. In *Vagueness in Communication*, pages 127–150. Springer, Berlin/ Heidelberg.

Latecki, L. J., Lakaemper, R., and Eckhardt, U. (2000). Shape descriptors for non-rigid shapes with a single closed contour. In *Proceedings of CVPR*, page 424–429.

Latecki, L. J., Lakaemper, R., and Wolter, D. (2005). Optimal partial shape similarity. *Image and Vision Computing*, 23(2):227–236.

Latecki, L. J. and Lakämper, R. (2000). Shape similarity measure based on correspondence of visual parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1185–1190.

Latecki, L. J., Lu, C., Sobel, M., and Bai, X. (2008). Multiscale random fields with application to contour grouping. In *Proceedings of Neural Information Processing Systems Conference (NIPS)*, pages 913–920.

Lawry, J. and Tang, Y. (2009). Uncertainty modelling for vague concepts: A prototype theory approach. *Artificial Intelligence*, 173(18):1539–1558.

Levinson, S. C. (1996). Frames of reference and Molyneux's question: Cross-linguistic evidence. In Bloom, P., Peterson, M., Nadel, L., and Garrett, M., editors, *Language and Space*, pages 109–169. MIT Press, Cambridge, MA.

Lewis, D. (1970). General semantics. *Synthese*, 22(1):18–67.

Ling, H. and Jacobs, D. W. (2007). Shape classification using the inner-distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:286–299.

Logan, G. D. (1994). Spatial attention and the apprehension of spatial relations. *Journal of Experimental Psychology: Human Perception and Performance*, 20(5):1015–1036.

Logan, G. D. (1995). Linguistic and conceptual control of visual spatial attention. *Cognitive Psychology*, 28(2):103–174.

Logan, G. D. and Sadler, D. D. (1996). A computational analysis of the apprehension of spatial relations. In Bloom, P., Peterson, M., Nadell, L.,

and Garrett, M., editors, *Language and Space*, pages 493–529. MIT Press, Cambridge, MA.

Lotto, R. B. and Purves, D. (1999). The effects of color on brightness. *Nature Neuroscience*, 2(11):1010–1014.

Mangold, R. and Pobel, R. (1988). Informativeness and instrumentality in referential communication. *Journal of Language and Social Psychology*, 7(3-4):181–191.

Marr, D. and Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London B: Biological Sciences*, 200(1140):269–294.

Mast, V. and Bergmann, E. (2013). Is it really that simple? the complexity of object descriptions in human-computer interaction. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Cognitive Science Society. to appear.

Mast, V., Couto Vale, D., Falomir, Z., and Elahi, M. F. (2014a). Referential grounding for situated human-robot communication. In Rieser, V. and Muller, P., editors, *Proceedings of SemDial 2014 – DialWatt*, pages 223–225. Association for Computational Linguistics.

Mast, V., Falomir, Z., and Wolter, D. (2016). Probabilistic reference and grounding with pragr for dialogues with robots. *Journal of Experimental and Theoretical Artificial Intelligence*.

Mast, V. and Wolter, D. (2013). Context and vagueness in REG. In *Proceedings of Production of Referring Expressions: Bridging the Gap betweeen Cognitive and Computational Approaches to Reference (PRE-CogSci 2013)*, Berlin.

Mast, V., Wolter, D., Klippel, A., Wallgrün, J. O., and Tenbrink, T. (2014b). Boundaries and prototypes in categorizing direction. In *Spatial Cognition IX*, pages 92–107. Springer, Berlin/ Heidelberg.

McMahan, B. and Stone, M. (2015). A Bayesian model of grounded color semantics. *Transactions of the Association for Computational Linguistics*, 3:103–115.

McTear, M. (2004). *Spoken Dialogue Technology: Toward the Conversational User Interface*. Springer, London.

Medin, D. L., Altom, M. W., and Murphy, T. D. (1984). Given versus induced category representations: Use of prototype and exemplar information in classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(3):333–352.

Mellish, C., Scott, D., Cahill, L., Paiva, D., Evans, R., and Reape, M. (2006). A reference architecture for Natural Language Generation systems. *Natural Language Engineering*, 12:1–34.

Menegaz, G., Troter, A. L., Sequeira, J., and Boi, J. M. (2007). A discrete model for color naming. *Journal on Applied Signal Processing*, 2007(1).

Meo, T., McMahan, B., and Stone, M. (2014). Generating and resolving vague color references. In Rieser, V. and Muller, P., editors, *Proceedings of SemDial 2014 – DialWatt*, pages 107–115. Association for Computational Linguistics.

Mervis, C. B. and Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, 32(1):89–115.

Miller, G. A. and Johnson-Laird, P. N. (1976). *Language and Perception*. Belknap Press.

Moratz, R. and Tenbrink, T. (2006). Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations. *Spatial Cognition and Computation*, 6(1):63–106.

Mori, G., Belongie, S., and Malik, J. (2001). Shape contexts enable efficient retrieval of similar shapes. In *Computer Vision and Pattern Recognition*, pages 723–730. IEEE Computer Society.

Myerson, R. B. (1991). *Game Theory: Analysis of Conflict*. Harvard University Press, Cambridge, MA.

Olson, D. R. and Ford, W. (1975). The elaboration of the noun phrase in children's description of objects. *Journal of Experimental Child Psychology*, 19:371–382.

Paek, T. and Horvitz, E. (1999). Uncertainty, utility, and misunderstanding: A decision-theoretic perspective on grounding in conversational systems. In *AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems, Cape Cod, MA*.

Palmer, S. (1999). *Vision Science: Photons to Phenomenology*. MIT Press, Cambridge, MA.

Palmer, S. E. and Schloss, K. B. (2010). An ecological valence theory of human color preference. *Proceedings of the National Academy of Sciences*, 107(19):8877–8882.

Passonneau, R. J. (1996). Using centering to relax Gricean informational constraints on discourse anaphoric noun phrases. *Language and Speech*, 39(2-3):229–264.

Pechmann, T. (1984). *Überspezifizierung und Betonung in referentieller Kommunikation*. PhD thesis, Universität Mannheim.

Pechmann, T. (1987). *Effects of Incremental Speech Production on the Syntax and Content of Noun Phrases*, volume 120 of *Arbeiten der Fachrichtung Psychologie der Universität des Saarlandes*.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, 27(1):89–110.

Pickering, M. J. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–189.

Plumert, J. M., Carswell, C., De Vet, K., and Ihrig, D. (1995). The content and organization of communication about object locations. *Journal of Memory and Language*, 34(4):477–498.

Pobel, R., Grosser, C., Mangold, R., and Hermann, T. (1988). Zum Einfluß hörerseitiger Wahrnehmungsbedingungen auf die Überspezifikation von Objektbenennungen. Technical report, Forschergruppe "Sprechen und Sprachverstehen im sozialen Kontext", Universität Mannheim.

Posner, M. I. and Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, 77(3p1):353–363.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, 3(3):382–407.

Regier, T. and Carlson, L. A. (2001). Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology: General*, 130(2):273.

Reiter, E. (1990). The computational complexity of avoiding conversational implicatures. In *Proceedings of the 28th annual meeting of the Association for Computational Linguistics*, pages 97–104. Association for Computational Linguistics.

Reiter, E. (1994). Has a consensus NL Generation architecture appeared, and is it psycholinguistically plausible? In *Proceedings of the Seventh International Natural Language Generation Workshop (INLG)*, pages 163–170, Kennebunkport, Maine.

Reiter, E. and Dale, R. (1992). A fast algorithm for the generation of referring expressions. In *Proceedings of the 14th conference on Computational linguistics*, pages 232–238. Association for Computational Linguistics.

Roorda, A. and Williams, D. R. (1999). The arrangement of the three cone classes in the living human eye. *Nature*, 397(6719):520–522.

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, 4(3):328–350.

Ross, R. J. (2009). *Situated Dialogue Systems*. PhD thesis, University of Bremen.

Ross, R. J. and Bateman, J. A. (2009). Daisie: Information state dialogues for situated systems. In Matouček, V. and Mautner, P., editors, *Text, Speech and Dialogue*, LNCS, pages 379–386. Springer.

Roy, D. K. (2002). Learning visually-grounded words and syntax for a scene description task. *Computer Speech and Language*, 16(3-4):353–385.

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

Russell, S. and Norvig, P. (2002). *Artificial Intelligence: A Modern Approach*. Prentice Hall.

Sarifuddin, M. and Missaoui, R. (2005). A new perceptually uniform color space with associated color similarity measure for content-based image and video retrieval. In *Proceedings of ACM SIGIR 2005 Workshop on Multimedia Information Retrieval (MMIR)*, pages 1–8.

Schober, M. F. (1993). Spatial perspective-taking in conversation. *Cognition*, 47(1):1 – 24.

Seaborn, M., Hepplewhite, L., and Stonham, T. J. (2005). Fuzzy colour category map for the measurement of colour similarity and dissimilarity. *Pattern Recognition*, 38(2):165–177.

Searle, J. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge, UK.

Sebastian, T. B., Klein, P. N., and Kimia, B. B. (2002). Shock-based indexing into large shape databases. In Heyden, A., Sparr, G., Nielsen, M., and Johansen, P., editors, *Computer Vision — ECCV 2002*, volume 2352 of *LNCS*, pages 731–746. Springer, Berlin/ Heidelberg.

Shi, H., Jian, C., and Rachuy, C. (2011). Evaluation of a unified dialogue model for human-computer interaction. *International Journal of Computational Linguistics and Applications*, 2(1-2):155–173.

Siddiqi, K., Shokoufandeh, A., Dickinson, S. J., and Zucker, S. W. (1999). Shock graphs and shape matching. *International Journal of Computer Vision*, 35(1):13–32.

Sivik, L. and Taft, C. (1994). Color naming: A mapping in the IMCS of common color terms. *Scandinavian Journal of Psychology*, 35(2):144–164.

Sonnenschein, S. (1984). The effects of redundant communications on listeners: Why different types may have different effects. *Journal of Psycholinguistic Research*, 13:147–166.

Soto-Hidalgo, J. M., Chamorro-Martínez, J., and Sanchez, D. (2010). A new approach for defining a fuzzy color space. In *Proceedings of 2010 IEEE International Conference on Fuzzy Systems (FUZZ)*, pages 1–6.

Spranger, M. (2011). *The Evolution of Grounded Spatial Language*. PhD thesis, Vrije Universiteit Brussels (VUB), Brussels, Belgium.

Spranger, M. and Pauw, S. (2012). Dealing with perceptual deviation - Vague semantics for spatial language and quantification. In Steels, L. and Hild, M., editors, *Language Grounding in Robots*, pages 173–192. Springer, Berlin/ Heidelberg.

Stamenković, D. (2011). Verbs and prototype theory: State of the art and possibilities. *English Studies Today: Views and Voices – Selected Papers from the First International Conference on English Studies – English Language and Anglophone Literatures Today (ELALT)*, pages 175–186.

Steedman, M. and Baldridge, J. (2011). Combinatory categorial grammar. *Non-Transformational Syntax: Formal and Explicit Models of Grammar. Wiley-Blackwell*.

Steels, L. (2008). The symbol grounding problem has been solved, so what's next? In De Vega, M., Glenberg, A., and Graesser, A., editors, *Symbols and Embodiment: Debates on Meaning and Cognition*, pages 223–244. Oxford University Press, Oxford.

Steels, L., Belpaeme, T., et al. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, 28(4):469–488.

Stoia, L., Shockley, D. M., Byron, D. K., and Fosler-Lussier, E. (2006). Noun phrase generation for situated dialogs. In *Proceedings of the Fourth International Conference on Natural Language Generation (INLG)*, pages 81–88. Association for Computational Linguistics.

Stone, M. (1998). *Modality in dialogue: planning, pragmatics and computation*. PhD thesis, Department of Computer and Information Science, University of Pennsylvania.

Stone, M. (2000). On identifying sets. In *Proceedings of the first International Conference on Natural Language Generation (INLG)*, pages 116–123. Association for Computational Linguistics.

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

Striegnitz, K., Denis, A., Gargett, A., Garoufi, K., Koller, A., and Theune, M. (2011). Report on the second second challenge on generating instructions in virtual environments (GIVE-2.5). In *Proceedings of the 13th European Workshop on Natural Language Generation (ENLG)*, Nancy, France. Association for Computational Linguistics.

Strohner, H., Sichelschmidt, L., Duwe, I., and Kessler, K. (2000). Discourse focus and conceptual relations in resolving referential ambiguity. *Journal of Psycholinguistic Research*, 29(5):497–516.

Super, B. (2004). Fast correspondence-based system for shape retrieval. *Pattern Recognition Letters*, 25(2):217–225.

Talmy, L. (1983). How language structures space. In Pick, H. and Acredolo, L., editors, *Spatial Orientation: Theory, Research, and Application*.

Talmy, L. (2000). *Toward a Cognitive Semantics*. MIT Press, Cambridge, MA.

Tellex, S., Knepper, R., Li, A., Rus, D., and Roy, N. (2014). Asking for help using inverse semantics. *Proceedings of Robotics: Science and Systems, Berkeley, USA*, 2.

Tellex, S., Kollar, T., Dickerson, S., Walter, M., Banerjee, A., Teller, S., and Roy, N. (2011). Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of AAAI Conference on Artificial Intelligence*.

Tenbrink, T. (2005). Identifying objects on the basis of spatial contrast: An empirical study. In Freksa, C., Knauff, M., Krieg-Brückner, B., Nebel, B., and Barkowsky, T., editors, *Spatial Cognition IV. Reasoning, Action, Interaction*, pages 124–146. Springer.

Tetreault, J. and Allen, J. (2004). Semantics, dialogue, and reference resolution. In *CATALOG '04*, Barcelona.

Theune, M., Touset, P., Viethen, J., and Krahmer, E. (2007). Cost-based attribute selection for GRE (GRAPH-SC/GRAPH-FP). In Belz, A. and Varges, S., editors, *Proceedings of the MT Summit XI Workshop on Using Corpora for NLG: Language Generation and Machine Translation (UCNLG+MT)*, pages 95–97. European Association for Machine Translation.

Tulving, E. (1962). Subjective organization in free recall of "unrelated" words. *Psychological Review*, 69(4):344.

van Deemter, K. (2002). Generating referring expressions: Boolean extensions of the incremental algorithm. *Computational Linguistics*, 28(1):37–52.

van Deemter, K. (2006). Generating referring expressions that involve gradable properties. *Computational Linguistics*, 32(2):195–222.

van Deemter, K. (2010). *Not Exactly: In Praise of Vagueness*. Oxford University Press, Oxford.

van Deemter, K. (2016). *Computational Models of Referring: A Study in Cognitive Science*. MIT Press.

van Deemter, K., Gatt, A., van der Sluis, I., and Power, R. (2012a). Generation of referring expressions: Assessing the incremental algorithm. *Cognitive Science*, 36(5):799–836.

van Deemter, K., Gatt, A., van Gompel, R. P., and Krahmer, E. (2012b). Toward a computational psycholinguistics of reference production. *Topics in Cognitive Science*, 4(2):166–183.

van Deemter, K., van der Sluis, I., and Gatt, A. (2006). Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 130–132. Association for Computational Linguistics.

van der Sluis, I. and Krahmer, E. (2004). The influence of target size and distance on the production of speech and gesture in multimodal referring expressions. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP'04)*, Jeju, Korea.

van Gompel, R., Gatt, A., Krahmer, E., and van Deemter, K. (2014). Overspecification in reference: modelling size contrast effects. In *Proceedings of the Architectures and Mechanisms for Language Processing (AMLaP) Conference [abstract]*, Edinburgh.

van Rooij, R. (2011). Vagueness and linguistics. In Ronzitti, G., editor, *Vagueness: A Guide*, pages 123–170. Springer Netherlands.

Viethen, H. A. E. (2011). *The Generation of Natural Descriptions: Corpus-Based Investigations of Referring Expressions in Visual Domains*. PhD thesis, Macquarie University, Sydney, Australia.

Viethen, J. and Dale, R. (2006). Algorithms for generating referring expressions: Do they do what people do? In *Proceedings of the Fourth International Natural Language Generation Conference (INLG)*, pages 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

Viethen, J. and Dale, R. (2008). The use of spatial relations in referring expression generation. In *Proceedings of the Fifth International Natural Language Generation Conference (INLG)*, pages 59–67. Association for Computational Linguistics.

Viethen, J., Goudbeek, M., and Krahmer, E. (2012). The impact of colour difference and colour codability on reference production. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society (CogSci)*.

Von Stechow, A. (1984). Comparing semantic theories of comparison. *Journal of Semantics*, 3(1):1–77.

Vorwerg, C. and Tenbrink, T. (2007). Discourse factors influencing spatial descriptions in English and German. In Barkowsky, T., Knauff, M., Ligozat, G., and Montello, D., editors, *Spatial Cognition V. Reasoning, Action, Interaction*, volume 4387 of *LNCS*, pages 470–488. Springer, Berlin/ Heidelberg.

Wang, N., Ai, H., and Lao, S. (2011). Computer vision – accv 2010: 10th asian conference on computer vision, revised selected papers, part iii. pages 171–184, Berlin/ Heidelberg. Springer.

Wilkes-Gibbs, D. and Clark, H. H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language*, 31(2):183–194.

Wittgenstein, L. (1953). *Philosophical Investigations*. Blackwell, Oxford.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8(3):338–353.

Zender, H., Kruijff, G.-J. M., and Kruijff-Korbayová, I. (2009). Situated resolution and generation of spatial referring expressions for robotic assistants. In *Proceedings of the 21st International Jont Conference on Artifical Intelligence (IJCAI)*, pages 1604–1609. Morgan Kaufmann Publishers Inc.

Zhekova, D. (2013). *Towards Multilingual Coreference Resolution*. PhD thesis, University of Bremen.

Zimmer, H., Speiser, H., Baus, J., Blocher, A., and Stopp, E. (1998). The use of locative expressions in dependence of the spatial relation between target and reference object in two-dimensional layouts. In Freksa, C., Habel, C., and Wender, K., editors, *Spatial Cognition: An Inderdisciplinary Approach to Representing and Processing Spatial Knowledge*, volume 1404 of *LNCS*, pages 223–240. Springer, Berlin/ Heidelberg.

Appendix A

Experiment Materials

A.1 Participant Instruction Reference Interpretation

Im Folgenden wirst du jeweils eine Objektbeschreibung, und dann eine Szene mit Objekten sehen. Deine Aufgabe ist es, die Objektbeschreibung zu lesen, und anschließend das Objekt in der Szene anzuklicken, das deiner Meinung nach beschrieben wurde. Beim Lesen der Objektbeschreibung, und beim Identifizieren des Objekts solltest du versuchen **möglichst schnell** zu sein.

- Vor der Beschreibung erscheint ein großes + auf dem Bildschirm.
- Schaue auf das +, bis die Objektbeschreibung erscheint.
- Lies dann die Beschreibung **möglichst zügig**. Klicke mit der linken Maustaste an einer beliebigen Stelle auf den Bildschirm, sobald du die Beschreibung verstanden hast.
- Wenn du sehr lange zum Lesen brauchst, geht das Experiment automatisch weiter. Das ist nicht schlimm, mach einfach weiter so gut du kannst.
- Es wird dann ein weiteres + erscheinen.
- Schaue auf das +, bis das Szenenbild erscheint.

- Wenn das Szenenbild erscheint, befindet sich der Mauszeiger in der Mitte der Szene.
- Klicke dann **so schnell wie möglich** mit der Maus auf das Objekt, das deiner Meinung nach beschrieben wurde.
- Wenn du unsicher bist, oder die Beschreibung auf mehrere Objekte zutrifft, klicke das Objekt an, das deiner Meinung nach am wahrscheinlichsten gemeint ist.
- Wenn du sehr lange brauchst um ein Objekt auszuwählen, geht das Experiment automatisch weiter. Das ist nicht schlimm, mach einfach mit der nächsten Aufgabe weiter.

Wichtig: Es gibt kein richtig oder falsch. Entscheidend dafür welches Objekt du anklickst ist, wie du die Beschreibung verstanden hast!

A.2 Participant Instruction Evaluating Referring Expressions

Im Folgenden wirst du jeweils eine Objektbeschreibung, und eine Szene mit Objekten sehen. In der Szene mit Objekten ist das beschriebene Objekt mit einem roten oder schwarzen Kreuz markiert. Deine Aufgabe ist es, die Objektbeschreibung zu lesen, und zu beurteilen, wie zutreffend und wie eindeutig die Beschreibung ist. Dazu kannst du jeweils einen Schieberegler zwischen den beiden Polen "trifft gar nicht zu" und "trifft voll und ganz zu" verstellen.

Für diese Aufgabe kannst du dir **so viel Zeit nehmen wie du willst**. Wichtig ist, dass du die Bewertung gewissenhaft vornimmst.

Wichtig: Es gibt kein richtig oder falsch, entscheidend für deine Bewertung ist, wie du die Beschreibung beurteilst!

A.3 Participant Instruction Producing Referring Expressions

Im Folgenden wirst du jeweils eine Szene mit Objekten sehen. Dabei wird ein Objekt mit einem schwarzen Pfeil markiert sein. Deine Aufgabe ist es, dieses Objekt zu beschreiben. Eine andere Person, die diese Szene ohne den Markierungspfeil sieht, sollte das markierte Objekt nur anhand deiner Beschreibung identifizieren können.

- Schaue dir die Szene an, und gib deine Beschreibung in das Textfeld ein.
- Wenn du fertig bist, klicke auf 'Weiter' um zur nächsten Szene zu gelangen.

Wichtig: Es gibt kein richtig oder falsch. Entscheidend dafür wie du das Objekt beschreibst, ist deine persönliche Auffassung davon, was eine geeignete Beschreibung ist!